DOCUMENT RESUME

ED 471 083 TM 034 589

AUTHOR	Fitzpatrick, Anne R.
TITLE	Investigating the Comparability of School Scores across Test Forms that Are Not Parallel. Technical Guidelines for Performance Assessment.
INSTITUTION	Council of Chief State School Officers, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2002-00-00
NOTE	50p.; Produced by the State Collaborative on Assessment and Student Standards.
CONTRACT	R279A50006
AVAILABLE FROM	Council of Chief State School Officers, One Massachusetts Avenue, Suite 700, Washington, DC 20001-1431 ($10). Tel: 202-408-5505; Fax: 202-408-8072. For full text: http://publications.ccsso.org
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Comparative Analysis; Mathematics Tests; Multiple Choice Tests; *Performance Based Assessment; *Scores; State Programs; *Test Format; Testing Programs
IDENTIFIERS	*Parallel Test Forms

ABSTRACT

This study, one of a series designed to answer practical questions about performance based assessment, examined the comparability of school scores on short, nonparallel test forms. The data were obtained from mathematics tests with both multiple choice (MC) and performance assessment (PA) items. The tests were administered in a statewide testing program in grades 4 and 8, with three test forms administered in each grade. Scores based on the PA items in the three forms were equated, and the comparability of these scores at the school level was investigated. The most important practical implication of the study's findings is that school-by-form interactions should be taken into account if the school scores by form are to be used. Findings indicate that even when forms are equated, if they are short and not parallel, school scores can vary substantially across the forms. Making forms parallel and long would reduce this problem considerably. Other practical implications are discussed, including the conclusion that any estimate of possible variation in school scores is better than no estimate at all. An appendix shows the calculation of statistics used in the analyses. (Contains 18 tables, 17 figures, and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 471 083

SCASS
State Collaborative on Assessment and Student Standards

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Buterbaugh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Investigating the Comparability of School Scores Across Test Forms That Are Not Parallel

Anne R. Fitzpatrick

TM034589

SCASS TGPA
Technical Guidelines for Performance Assessment

A Publication of the Council of Chief State School Officers

BEST COPY AVAILABLE

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues. Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA) consortium. The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

G. Thomas Houlihan
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards

Copyright © 2002 by the Council of Chief State School Officers, Washington, DC

SCASS TGPA
Technical Guidelines for Performance Assessment

Investigating the Comparability of School Scores Across Test Forms That Are Not Parallel

Anne R. Fitzpatrick
CTB/McGraw-Hill

Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, DC 20001
202-408-5505

This report was prepared for submission under contract with the Council of Chief State School Officers. The preparation of the report was funded by OERI Grant No. R279A50006. The views and opinions expressed in this report are not necessarily those of the United States Department of Education, the Council of Chief State School Officers, the State Collaborative on Assessment and Student Standards, or the states participating in the study.

ACKNOWLEDGMENTS

This project was conducted with the assistance of members of CCSSO's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA). These individuals helped in developing the research design. They also contributed data and reviewed draft versions of this report. The authors wish to thank the following people for these efforts:

Jonathan Dings	Kentucky Department of Education*
James Friedebach	Missouri Department of Elementary and Secondary Education
Linda Hansche Despriet	Georgia Assessment Project*
James Hertzog	Pennsylvania Department of Education*
Duncan MacQuarrie	Washington Department of Education*
James Masters	Pennsylvania Department of Education
Jeffrey Nellhaus	Massachusetts Department of Education
Ellie Sanford	North Carolina Department of Public Instruction*
Gloria Turner	Alabama State Department of Education
Phoebe Winter	Council of Chief State School Officers*

* At time of development.

EXECUTIVE SUMMARY

This report presents one of a collection of studies sponsored by the Technical Guidelines in Performance Assessment consortium of the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards. The series of studies is designed to answer practical questions about the technical properties of assessment instruments that include constructed-response item formats and to provide information about how states are using performance assessments.

The current research examined the comparability of school scores on short, non-parallel test forms. The data were obtained from mathematics tests having both multiple-choice (MC) and performance assessment (PA) items. The tests were administered in a statewide testing program in grades 4 and 8. In each grade, three forms of the mathematics test were administered. The scores based on the PA items in the three forms were equated, and the comparability of these scores at the school level was investigated.

The most important practical implication of the study's findings is that school-by-form interactions should be taken into account if school scores by form are to be used. The findings indicate that even when forms are equated, if they are short and not parallel, school scores can vary substantially across the forms. Equating does not fix this kind of variation. Making forms parallel and long would reduce this problem considerably. The findings include other practical implications:

1. School-by-form interactions, that is, variation in school scores across test forms, can occur when the test forms are not parallel in terms of their content and their psychometric properties.

2. When the test forms are not parallel and are very short, as they were in this study, the interactions for some schools can be substantial, producing sizable differences in school scores across forms.

3. When multiple test forms that are not parallel must be administered in a given year, it is prudent to administer all forms within a school and obtain school-level results by aggregating individual student scores across forms.

4. When school scores by form are to be used, an estimate of the degree to which these scores will vary due to school-by-form interactions should be provided.

5. Any estimate of the possible variation in school scores is better than no estimate at all.

6. The magnitude of school-by-form interactions was found to vary by school size. Therefore, it seems prudent to estimate school-by-form interactions separately for groups of schools that differ in size.

These results provide state assessment staff with information about the characteristics of tests that are needed to provide comparable school-level results—information useful in designing or remodeling their assessment programs. The full report describes techniques that state assessment developers can use to explore and evaluate the quality of their assessment data.

CONTENTS

INTRODUCTION 1

METHOD 3

Tests Administered 3
Analyses 3

RESULTS/DISCUSSION 7

Test Characteristics 7
Findings Based on Scale Scores 8
Analyses of All Schools within a Grade 8
Analyses by School Size Categories within a Grade 8
Findings Based on PAACs 9

PRACTICAL IMPLICATIONS: *The Importance of Building Confidence Bands for School Scores* 13

CONCLUSIONS 15

REFERENCES 17

APPENDIX 19

TABLES 21

Table 1: Skills Measured by the PA Items in Each Form 21
Table 2: School Size Categories Used in Each Grade 22
Table 3: Descriptive Raw Score Test Statistics 22
Table 4: Results of Scaling MC and PA Items 22
Table 5: Scale Score Statistics Describing Student-Level Performance for All Students in the State 23
Table 6: Statistics Describing School-Level Scale Scores for All Schools* 23
Table 7: Statistics Describing School-Level Scale Scores by School Size* 24
Table 8: Mean Scale Scores by Form for the Grade 4 Schools with the Largest SDM_ys, by School Size Category 25
Table 9: Mean Scale Scores by Form for the Grade 8 Schools with the Largest SDM_ys, by School Size Category 25
Table 10: Mean Scale Scores by Form for the Grade 4 Schools with the Smallest SDM_ys, by School Size Category 26
Table 11: Mean Scale Scores by Form for the Grade 8 Schools with the Smallest SDM_ys, by School Size Category 26
Table 12: Student PAACs by Form for All Students in the State 27
Table 13: Statistics Describing School PAACs for All Schools* 27
Table 14: Statistics Describing School PAACs by School Size Category* 28
Table 15: PAACs by Form for the Grade 4 Schools with the Largest SDP_ys by School Size Category 29
Table 16: PAACs by Form for the Grade 8 Schools with the Largest SDP_ys by School Size Category 29
Table 17: Calculation of Standard Errors and Confidence Intervals for School Mean Scale Scores 30
Table 18: Calculation of Standard Errors and Confidence Intervals for School PAACs 30

FIGURES 31

Figure 1: Frequency distributions of the grade 4 equated scores for Forms A, B, and C 31
Figure 2: Frequency Distributions of the Grade 8 Equated Scores for Forms A, B, and C 32
Figure 3: Cumulative frequency distribution of grade 4 equated scores for Forms A, B, and C 33
Figure 4: Cumulative Frequency Distribution of the Grade 8 Equated Scores for Forms A, B, and C 34
Figure 5a: Standard error curves for the grade 4 equated Forms A, B, and C 35
Figure 5b: Standard error curves for the grade 8 equated Forms A, B, and C 35
Figure 6a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 4 small school category 36

BEST COPY AVAILABLE

Figure 6b: Mean scale scores by form for the five schools with the smallest SDM_ys in the grade 4 small school category 36
Figure 7a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 4 small school category 37
Figure 7b: Mean scale scores by form for the five schools with the smallest SDM_ys in the grade 4 medium-sized school category 37
Figure 8a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 4 lage school category 38
Figure 8b: Mean scale scores by form for the five schools with the smallest SDM_ys in the Grade 4 Large School category 38
Figure 9a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 8 small school category 39
Figure 9b: Mean scale scores by form for the five schools with the smallest SDM_ys in the grade 8 small school category 39
Figure 10a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 8 medium-sized category 40
Figure 10b: Mean scale scores by form for the five schools with the smallest SDM_ys in the grade 8 medium-sized school category 40
Figure 11a: Mean scale scores by form for the five schools with the largest SDM_ys in the grade 8 small school category 41
Figure 11b: Mean scale scores by form for the five schools with the smallest SDM_ys in the grade 8 large school category 41
Figure 12: PAACs by form for the five grade 4 schools with the largest SDP_ys 42
Figure 13: PAACs by form for the five grade 8 schools with the largest SDP_ys 43
Figure 14: Percentage frequency distributions of SDM_ys for grade 4 by school size category 44
Figure 15: Percentage frequency distributions of SDM_ys for grade 8 by school size category 45
Figure 16: Percentage frequency distributions of SDP_ys for grade 4 by school size category 46
Figure 17: Percentage frequency distributions of SDP_ys for grade 8 by school size category 47

BEST COPY AVAILABLE

INTRODUCTION

In this study we examined the comparability of school-level scores based on test forms that contained performance assessment (PA) items and had been equated. More specifically, we investigated the degree to which the school-level scores varied across test forms.

The test forms analyzed differed somewhat in terms of the mathematics skills that they measured and in terms of their psychometric properties. In the testing program in which these forms were used, performance at the school level was determined by aggregating student results over test forms before making comparisons between schools within or across test years.

The current study demonstrates why this kind of aggregation across forms is highly desirable and appropriate when test forms are not parallel. We found that school scores can vary considerably across individual test forms when these forms have different content and measurement properties. This occurred despite the fact that we had equated the forms. Were one to assess a school's performance using just one of these non-parallel test forms, the conclusions one might draw could be in error because of the interaction between school and form.

Variation in school scores across forms has been called school-by-form interaction (Cronbach, Bradburn, & Horvitz, 1994; Cronbach, Linn, Brennan, & Haertel, 1997; Yen, 1997a). Cronbach et al. (1994) considered these interactions a component of measurement error.

School performance can vary over non-parallel test forms for a variety of reasons. Curriculum often varies across schools, and a given school or teacher may happen to give more or less emphasis to the skills that happen to appear on one form. In addition, students in different schools may have varying opportunities to learn the material covered by a particular test form. There are also school environment variables that could cause schools' performance to vary over test forms.

In several recent generalizability studies, school-by-form interactions have been found to be very large and an important factor to include in an estimate of school-level standard errors (Cronbach et al., 1994, 1997; Yen, 1997a, b). Yen (1997a), for example, found that school-by-form interactions across mathematics forms accounted for up to 91% of the squared standard error estimates for very large schools.

Interaction effects when PAs are administered have been found at the student level as well. Students' performance has been commonly found to vary over tasks (e.g., Lane, Stone, Ankenmann, & Liu, 1993; Shavelson, Baxter, & Gao, 1993).

In the current study, we assess school-by-form interactions both in terms of scale scores developed using item response theory (IRT) and in terms of the percentage of students at or above a cutscore (PAACs). Both scale scores and PAACs are widely used to describe student achievement at the school level. Therefore, we anticipated that our research was likely to be most useful to practitioners if findings in both metrics were presented.

METHOD

TESTS ADMINISTERED

Items from an annual, statewide standards-based mathematics assessment in grades 4 and 8 were used in the current study. In each test form, 80 multiple-choice (MC) items were used to measure students' mathematics computation and application skills. In addition, three or four performance assessment (PA) items, each scored on a four-point (0, 1, 2, and 3) scale, were used to measure higher levels of thinking involving these skills. Each of these items took students about 10 minutes to complete. The content for both the MC and PA items was sampled from a domain of content defined by the state's curriculum guide.

Three test forms were developed for each grade, and the three forms were randomly assigned to students within a school. Thus, the groups of students who took the three forms administered in each grade were randomly equivalent.

ANALYSES

The three sets of PA items administered in each grade were the focus of the current study. We called these sets of items Forms A, B, and C. As noted previously, these item sets contained only three or four items, making each set very short. In addition, as Table 1 shows, these forms did not measure the same skills.

To obtain parameter estimates for the PA items, they were scaled together with the MC items that the students took. This was accomplished using a combination of item response theory (IRT) models and CTB's PARDUX Version 4.1 (Burket, 1996) scaling program. All the students in the state who took the form were used to calibrate the items in it.

The MC items were scaled using the three-parameter logistic (3PL) model. The 3PL model (Lord, 1980) defines the probability that an examinee with ability θ will correctly answer the jth item as

$$P_j(\theta) = P(X_j = 1|\theta) = C_j + \frac{(1 - C_j)}{1 + \exp(-1.7 A_j(\theta - B_j))},$$

where A_j, B_j, and C_j, respectively, refer to the discrimination, location, and "pseudo-guessing score level" parameters of the item.

The PA items were scaled using the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993). Given a PA item with m_j score levels assigned integer scores ranging from 0 to m_j -1, the 2PPC model states that the probability that an examinee with ability θ will perform at the k-th score level on item j can be expressed as

$$P_{jk}(\theta) = P(X_j = k - 1|\theta) = \frac{\exp(z_{jk})}{\sum_{i=1}^{m_j} \exp(z_{ji})}$$

where

$$z_{jk} = \alpha_j (k-1)\,\theta - \sum_{i=0}^{k-1} \gamma_{jk},$$

and $\gamma_{j0} = 0$. The 2PPC model is a special case of Bock's (1972) nominal model.

Fit was assessed using the Q_1 statistic described by Yen (1984) to evaluate the MC items. A generalization of this statistic was used to evaluate the fit of the multi-level PA items. As described by Yen, Q_1 is a Pearson chi-square of the form

$$Q_{1j} = \sum_{i=1}^{I} \frac{N_{ji}(O_{ji} - E_{ji})^2}{E_{ji}} + \sum_{i=1}^{I} \frac{N_{ji}[(1-O_{ji}) - (1-E_{ji})]^2}{1-E_{ji}}$$

Here, N_{jm} is the number of examinees in cell m for item j. O_{jm} and E_{jm} are, respectively, the observed and expected proportions of examinees in cell m who attained the maximum possible score on the item.

Because Q_1 and its generalization are affected by the number of score points associated with each item as well as by sample size, the statistic was transformed to a standardized z-statistic, where

$$z_j = \frac{Q_{1j} - DF}{(2DF)^{\frac{1}{2}}},$$

and DF refers to the number of degrees of freedom associated with an item. At CTB, our rule of thumb for sample sizes of 1,500 is to use $z \geq 4.00$ as the criterion for identifying items with possible poor fit. For sample sizes above 1,500, the criterion is adjusted by a factor equal to N/1500, where N refers to the sample size with which one is working. In our study, a criterion of $z \geq 67.00$ was used with our large sample sizes, which exceeded 25,000 cases.

To assess whether students' responses to the MC and PA items were sufficiently unidimensional to be scaled together using the combination of 3PL and 2PPC models, Yen's (1984) Q_3 statistic was calculated. Q_3 can serve as an indicator of the degree of local item dependence (LID) among test items.

The Q_3 statistic essentially involves the calculation of the correlation between examinees' item scores once their trait estimates have been removed. A substantial residual correlation between an item pair indicates that examinees' scores on these items are affected by something over and above that which has been captured by the unidimensional trait score.

To calculate the Q_3 statistic, define a difference score, d_{ij}, as

$$d_{ij} = \mu_{ij} - \hat{P}_i(\hat{\theta}_i)$$

where u_{ij} is the item score of the ith examinee on the jth item. The correlation between the difference scores for items h and j taken over all examinees then can be expressed as

$$Q_3 = r_{hj}.$$

As a rule of thumb, Q_3 values ≥ 0.200 have been used at CTB to flag items as locally dependent.

After the MC and PA items were scaled, student scores were estimated using only the three to four PA items that they took. A cumulative score distribution for each PA-only form was developed, and linear equipercentile procedures[1] were used to equate the PA scores students obtained on Forms B and C with the scores that students obtained on the target, which was arbitrarily selected to be Form A.

In both grades 4 and 8, the equated scale scores were reported on a scale that ranged from 300 to 700. The relationship between our score scales and those used by the state is not known. The score scales for the two grades were not linked vertically; rather, in both grades students' true thetas were assumed to be distributed normally with a mean of 500 and a standard deviation of 50.

Of primary interest in the current study was the question of the degree to which the equating relationships established using all of the students in the state would hold when scores at the school level were compared. To examine this question, we first looked at the consistency of performance across forms at the state level and compared that with the consistency of performance at the school level. In addition, to ascertain the degree to which the variability we observed across forms at the school level was due to stable school-by-form interactions rather than the unstable effects of sampling, we compared this variability with what would occur because of sampling effects.

To make these comparisons, we calculated several variances, described in general terms below. More details about how they were developed are given in Appendix A.

For the analyses involving scale scores, we first calculated the variance of students' scores on each form within a school and found the average variance across the three forms administered in each school. Then we summed the averages for all schools and found the average within form variance for the schools. The square root of this result was called the average standard deviation of student scores within schools (Average SDWS).

In addition, for each school we calculated the variance of its mean scores across the three forms. Then we summed the variances for all schools and found the average variance in mean scores for the schools. Its square root was called the average standard deviation of means (Average SDM) across forms. Finally, we used the Average SDWS, essentially a standard error (SE) term, to estimate how much a scale score mean on a form might vary simply due to sampling; we called this estimate SDM_s.

Because some research has suggested that the size of school-by-form interactions might vary with school size (see Yen, 1997a), we conducted our analyses both for all schools together and after grouping schools by size. Table 2 shows the size categories established for each grade. The school sizes used to define the category boundaries were somewhat arbitrarily decided, but resembled those used by Yen (1997a).

In grade 4, a cutscore was set at a scale score of 475, which corresponded to the 40th percentile on the target used for the equating. In grade 8, the cut score was set at a scale score of 461, which corresponded to the 45th percentile on the target form. A

[1] A linear equipercentile procedure involves carrying out an equipercentile equating and then finding the line that best fits the equipercentile results. The linear function that defines this line is used to transform both students' scores and the item parameters on which these scores are based to the target score scale.

cutscore nearer the 40th percentile was preferred, but there was not an obtainable scale score corresponding to this percentile. Cutscores near the 40th percentile were chosen as realistic examples of cutscores that might be set. As will be noted subsequently, the forms we analyzed also were very difficult; we had to choose relatively high cutscores just to include in our analyses students with scale scores above the lowest obtainable scores on the forms that we analyzed.

Analyses parallel to those involving the scale scores were carried out for the PAACs. The average standard deviation of PAACs across forms was called the Average SDP, and the estimate of how much a PAAC might vary due to sampling was called the SDP_s.

Our final analysis involved using the SDMs and SDPs to create confidence intervals that could be used to interpret school-level scale score means and PAACs. These confidence intervals provide one simple way of making clear to users of test results what level of precision they have. The SDMs and SDPs are useful because they reflect the combined effects of sampling error and the school-by-form interactions. We demonstrate how they can be used to convey to test users the level of uncertainty with which school-level results should be viewed.

RESULTS/DISCUSSION

TEST CHARACTERISTICS

Table 3 provides raw score statistics that describe the test forms analyzed in each grade. The observed percentage of maximum (OPM) refers to the mean raw score divided by the maximum possible score (MPS) on the form. The OPMs were in the 30% range at grade 4, indicating that the average student achieved about one-third of the maximum possible score. Thus, the grade 4 forms were difficult. With OPMs ranging between 13% and 22%, the grade 8 forms were extremely difficult.

The coefficient alphas for the test forms in both grades were low, most obviously because of test length; if the forms had been five times longer (i.e., 15 to 20 items), these coefficients would be in the 0.80s. It is likely that the difficulty of the test forms in both grades also contributed to the size of coefficient alphas by restricting the range of student scores.

Table 4 provides some information useful for evaluating the scaling. It shows that two or three items in each grade did not fit the models well; only one of these items was a PA item. No LID among the PA items was detected using the Q_3 statistics reported for each grade, suggesting that these items did not measure any skill over and above those measured by the other items with which they were scaled. For a few MC items in the grade 8 tests, no convergent parameter estimates were obtained.

Table 5 summarizes the student-level results in each grade in terms of the scale scores obtained after the forms had been equated. The summary statistics suggest that scores from the three forms were well aligned. As shown, the means and standard deviations of the scores for the three forms within each grade generally were very similar. The one exception was Form C in grade 4, where the standard deviation after equating remained larger than that observed for the other two grade 4 forms; this was due to the greater percentages of students with scores of 300 on this form.

In both grades, the standard deviations of the scores were sizable because of the large percentages of students whose scores did not exceed 300. Note that in grade 8, more than one quarter of the students did not score above 300.

Plots showing the scale score frequency distributions for each form are given in Figures 1 and 2 for grades 4 and 8, respectively. Note how discrete the score distributions were. Each distribution has distinct steps, and there are not many steps. This is because there were so few items and possible score points in each test form. Discrete score scales are typical of very short tests that have few items and/or score points.

When only a few scores are obtainable on a test form and these scores must be used to equate forms, some difficulties can arise. Consider as an example the cumulative frequency distributions shown in Figures 3 and 4. These plots show the grade 4 and grade 8 students' equated scores on Forms A, B, and C. Note in the plots how many instances there are where the steps in the cumulative distributions do not rise at the same time and/or to the same degree. If one picks a percentile and draws a horizontal line at that percentile through the three cumulative distributions, one can see what

scale score corresponds to the percentile on each form. After an ideal equating, the scale scores should be the same.

Figures 5a and 5b show plots of the standard error of measurement curves for each form in the two grades. They show that the forms were not equally precise throughout the score range. The grade 8 forms were particularly different in terms of their accuracy near the cutscore.

FINDINGS BASED ON SCALE SCORES

ANALYSES OF ALL SCHOOLS WITHIN A GRADE.

In Table 6, school-level scale score results are given. As was shown in Table 5, the within-school variance in student scores (Average SDWS) was much greater in grade 8 than in grade 4. This was because of the extreme difficulty of the grade 8 forms. The Average SDMs also were slightly larger in grade 8.

The SDM_ss decreased as grade level increased. This is predictable since the SDM_ss are standard error (SE) terms. All other things being equal, SEs will decrease as sample size increases; the grade 8 average school size was considerably larger than that in grade 4.

Note that the average standard deviation of form means (Average SDM) reported in Table 6 was about one-fourth the size of the average standard deviation of student scores (Average SDWS) within schools. This fraction is large and suggests that on the average, the schools' form means varied considerably.

We wanted to ascertain what portion of the variation in schools' form means reflected school-by-form interactions. To do this, we subtracted from the SDM calculated for each grade the SDM_s for the grade after squaring both terms.

The ratios in the far right-hand column of Table 6 show our results. In grade 4, 34% of the variance in the means reflected school-by-form interactions. At grade 8, fully 71% of the variance was attributable to these interactions.

ANALYSES BY SCHOOL SIZE CATEGORIES WITHIN A GRADE.

To investigate the effects of school size on comparability, we subsequently recalculated the SDMs and SDM_ss after assigning schools to size categories. The scale score results by the three school size categories that we used are given in Table 7.

In grade 4, the average SDMs decreased as school size increased. The average SDM for the large schools was about one-third less than that for the small schools. In grade 8, there was a decrease of about 40% in the Average SDM as school size increased from small to medium. However, the SDMs for the medium-sized and large schools were very close, indicating that on average, the variability in mean scores over forms did not differ much for the schools in these two categories. Table 7 shows that the SDM_ss decreased as school size increased, as would be expected.

The findings in the right-hand column indicate what percentage of the variability in form means was attributable to school-by-form interactions. In grade 4, this percentage varied from 22% and 36% over the size categories. It increased as school size increased. In grade 8, the percentage varied from 54% to 73%, again increasing as school size increased.

We wanted to get a clearer picture of what the variation in school means looked like. Accordingly, within each size category we sought out the five schools with the largest and smallest SDM_ys. By SDM_y we mean the standard deviation of the form means we calculated for each school y. (These values were subsequently averaged to find the SDMs.)

In Tables 8 and 9 are the mean scale scores for the five schools with the largest SDM_ys in grades 4 and 8, respectively. In Tables 9 and 10 are the means for the five grade 4 and five grade 8 schools that had the smallest SDM_ys. Plots of the findings are given in Figures 6 through 11.

Tables 8 and 9 show that in both grades the variation in form means for the schools with the largest SDM_ys was fairly impressive; in most instances, they were at least twice the size of the average SDM calculated for their school size category. In Table 9, look at the results for the first small school listed; its mean scale scores ranged from 397 to 660 over the three forms, a total range of 263 scale score points! Even the means for Small School #5 in Table 9 had a range of 168 scale score points. As school size increased, it is evident that the variability in means declined, but even for the five large schools listed in Table 9 the SDM_ys were at least 38 points! It is important to remember that some of the variation in form means is attributable to sampling error, but the data and plots give a feel for how much means can vary, regardless of the source of this variation. On the other hand, Tables 10 and 11 show schools with small SDM_ys, and they indicate that these schools had results that were very stable over forms.

Large school-by-form interactions like some of those described here certainly would argue against comparing schools that have taken different forms or comparing performance on different forms administered in different test years. Imagine that Forms A, B, and C were administered in sequential test years, say 1996, 1997, and 1998, and used for school accountability. Now say that you were the principal of School #3 in the small size category shown in Table 8. From 1996 to 1997, your grade 4 mean scores dropped by 117 score points, and from 1997 to 1998 they dropped another 49 score points, resulting in a total drop of 166 score points. How would it be possible to avoid the conclusion that something was terribly amiss in School #3?

Of course, in actuality it is possible that there was nothing amiss at the school. If the students who took the forms were equally able, the dramatic changes in the school's mean scale scores across forms were due to school-by-form interactions and sampling error.

In considering these results, it is important to remember that the forms we analyzed were not designed to be parallel measures of the same skills, and they were very short, consisting of only three or four items. As such, we are depicting a "pretty-bad-if-not-worst-case" scenario. Were test length increased substantially and the forms built to be comparable in terms of their content and psychometric properties, the school-by-form interactions would be much smaller than those we found.

Findings Based on PAACs

The PAACs by form for all grade 4 and grade 8 students in the state are given in Table 12. These PAACs were based on the cut scores of 475 and 461, respectively, used in the two grades. They were calculated by finding each student's scale score and then classifying each student as passing or failing in light of this score.

Since the forms were equated and the students who took the forms were randomly equivalent in ability, one might expect that at the state level the PAACs for all three forms within a grade would be the same. Table 12 reveals that this was not quite the case. In grade 4, the PAAC for Form C was three percentage points lower than the PAACs for Forms A and B. In grade 8, the PAACs for the three forms differed by up to 14 percentage points.

Recall the scale score summary statistics given in Table 5. As previously noted, except for the scores associated with Form C in grade 4, the scale score results suggested that at the state level the distributions of scores were very similar across forms. Yet, the data given in Table 12 show PAACs that were not the same across forms; this was especially true at grade 8.

Findings based on scale score and PAACs were not the same for several reasons. First, the scale score statistics and PAAC percentages capture different information about score distributions. Scale score means and standard deviations give a general picture of two characteristics of the entire score distributions, that is, where the scores are centered and how much the scores vary about this center on average. PAACs refer to results at a particular point in a score distribution, that is, the percentage of students who score at or above this point. These percentages can vary considerably over score distributions, even when the means and standard deviations of these distributions are the same.

A second reason is that PAACs reflect the reduction of a continuous score scale to a scale with only two scores, pass and fail. Information is lost when this simplification occurs: Students who achieve passing status are assigned the same dichotomous score, but they may differ in their skills and scale scores. Thus, through this loss of information dichotomization can make different things look similar or, indeed, similar things look different (as would be the case for students with scale scores just above and below the cutscore).[2]

In Table 13 are statistics describing the PAACs for all schools within each grade. The average PAACs by form for the schools resembled the values reported in Table 12, as might be expected.

It is noteworthy that the standard deviations (SDPs) of the school PAACs were similar in grades 4 and 8. This was not the case for the scale scores reported in Table 6, where the grade 8 SDMs were about 25% higher. Evidently, schools in the two grades differed less when evaluated in terms of their passing rates than they did when evaluated in terms of their scale score distributions. This makes sense. We decided that the grade 8 SDMs were large because of the extreme difficulty of the grade 8 forms; many students scored at the lowest obtainable scale score. The PAACs are based on students' pass/fail status only. They are not affected by how far above or below a cutscore the students' scores lie.

Comparing the values in the right-hand column of Table 13 with those reported on the basis of scale scores in Table 6, we see that the grade 4 ratios in the two tables were quite similar; they indicated that about 19% or 20% of the variance in the grade 4 PAACs was attributable to school-by-form interactions. In grade 8, however, more variation in the schools' PAACs was attributable to school-by-form interactions than that attributed when scale scores were considered. Evidently, the content differences

[2] The fact that different inferences might be drawn from results based on PAACs and scale scores suggests that equating results should be evaluated in both metrics when the comparability of forms in terms of PAACs, as well as scale scores, is important.

between forms affected grade 8 school pass rates more than these differences affected their scale scores.

PAAC results by school size are given in Table 14. The Average SDPs decreased as school size increased, and so did the SDP_ss. Turning to the ratios reported in the right-hand column, we see that the percentage of variance in the PAACs attributable to school-by-form interactions increased across grades.

Earlier data and plots were given for the five schools in each size category with the largest variation in their school means over forms. In Tables 15 and 16 we provide the corresponding data for the five schools in each size category with the largest variation in their PAACs over forms; it turned out that these schools were the same as those noted previously to have the highest SDM_ys.

Figure 12 gives plots of the grade 4 data given in Table 13. School #1 in the small school size category again had the most dramatic results. The range of PAACs for this school was 86 percentage points, changing over Forms A, B, and C from 72% to 86% to 0%, respectively. The ranges of PAACs for the medium and large grade 4 schools were substantially less than for the small schools, although their SDP_ys still were two to three times the average for their size group.

In grade 8, the schools with the largest SDP_ys similarly had two to three times the variation in their form PAACs than did the average for their size group. Plots for these schools are given in Figure 13. The range of PAACs over forms for the grade 8 schools was considerably greater than that observed for grade 4; this was also the case when scale scores were considered.

PRACTICAL IMPLICATIONS

THE IMPORTANCE OF BUILDING CONFIDENCE BANDS FOR SCHOOL SCORES

Our findings show that school scores can vary substantially across forms when the forms differ in terms of the skills they measure and their psychometric properties. *The most important practical implication of these findings is that school-by-form interactions should be taken into account if school scores by form are to be used.*

In this section we describe a relatively simple, empirical approach to building confidence bands that will enable more accurate reporting and use of school scores by form.[3] More technical and complex approaches are described by Cronbach et al. (1994, 1995) and Yen (1997a). These approaches are based on concepts of generalizability theory and analysis of variance methodology. The most appropriate method for estimating the sources of variation in school scores is still being debated, but as Yen (1997a) noted, it is far better to provide some estimates that might later be improved upon than to provide no estimates at all.

Our approach involves using just the SDMs and SDPs that we previously calculated. As noted earlier, the SDMs and SDPs reflect the effects of both sampling and school-by-form interactions. Thus, they capture the variation attributable to two of the most important variables that could affect school scores.

Tables 17 and 18 show the confidence bands for the schools that we analyzed in this study. We calculated values for all of the schools in each grade as well as separately for each school size group. The confidence bands can be generally expressed as

$$M_i = \pm z_a \sigma_w / \sqrt{F},$$

where z_a refers to the standard normal deviate with $\alpha/2$ of the distribution in each tail, and σ_w is estimated by either the SDM or the SDP. For Tables 17 and 18 we calculated a 90% confidence interval. To do this, z_a was set equal to 1.65.

Table 17 shows that in grade 4 the confidence intervals for school scale scores declined from 50 scale score points to 28 scale score points as school size increased. In grade 8 they declined from 64 scale score points for small schools to about 40 scale score points for the medium-sized and large schools.

Table 18 shows that in grade 4 the confidence intervals for the PAACs decreased from 26 to 16 percentage points as school size increased. In grade 8, they dropped from 30 percentage points for the small schools to about 20 points for the larger schools.[4]

When considering these findings, it is important to keep in mind that the SEs given in Tables 17 and 18 provide an estimate of the error in a single school mean or PAAC. If one wanted to compare the results from two schools or from two test years for a single school, each measure has a standard error; therefore, the comparison would involve two confidence intervals.

[3] Such confidence intervals are intended to assist in the interpretation of school scores. They should not be used for statistical tests.
[4] It is worth noting here that the Select Committee (Cronbach et al., 1994) stated that a confidence interval of about 8% was a reasonable target for the accuracy of a school passing rate.

It must be evident that the SEs we estimated using our very short, non-parallel forms were very large—so large that they permit few meaningful comparisons to be made. The SEs would be smaller if the forms were made more comparable, if more forms were spiraled within a school, and if the forms were longer. The passing rate also directly affects the sampling error of the PAACs (Yen, 1997a).

There are limitations in our simplified approach that are important to mention. Our sampling estimates were not adjusted to reflect assumptions made about the universe from which the student samples and test forms were drawn and about the method used to draw the samples and forms. Our estimates of school-by-form interactions resembled but were not identical to results that one would obtain using generalizabillity theory and analysis of variance procedures. Finally, our analyses by size category did not involve a balanced design. To accommodate this kind of design, a target school size for each category would be identified. Schools with enrollments lower than this size would be excluded from the analysis. In schools with enrollments greater than this target, students would be randomly excluded to produce samples equal to the target.

The advantage of the balanced design is that the findings are accurate for the schools analyzed. In our study, the findings for each size category were averages that invariably either over-estimated or under-estimated the effects for some schools. To see this, consider the plots in Figures 14 through 16. They show for each size group the frequency distributions of the SDM_ys and SDP_ys that were calculated. The Average SDM and Average SDP used to calculate the SDMs and SDPs used to compute our SE(SS)s and SE(PAAC)s are marked by a dotted line in each plot.

The plots show that the SDM_ys and SDP_ys for the schools in the small school size group varied tremendously. Indeed, as seen in Tables 8 and 10, grade 4 schools in this group could have SDMs as small as 1 point or as large as 115 points. One might think that given this variability it would be more accurate to calculate individual SDM_ys or SDP_ys for each school. However, these values are hugely unreliable due to sampling effects (Cronbach et al., 1994), and so average values based on groups of schools are required.

Our findings did vary somewhat by school size, especially in grade 4 and among small schools in both grades. This implies that some stratification of schools by size would be useful when SEs must be calculated. The variability in SDM_ys and SDP_ys evident in the plots for our small school category suggests that it might be beneficial to subdivide our small school group to obtain average SDMs and SDPs that would be more representative. If PAACs were the only statistic of interest, it also might be sensible to combine the schools in the medium-sized and large school categories, since the results and the distributions of SDP_ys for these two groups were quite similar.

CONCLUSIONS

The following conclusions can be drawn from our research:

1. School-by-form interactions, that is, variation in school scores across test forms, can occur when the test forms are not parallel in terms of their content and their psychometric properties.

2. When the test forms are not parallel and are very short, as they were in this study, the interactions for some schools can be substantial, producing sizable differences in school scores across forms.

3. When multiple test forms that are not parallel must be administered in a given year, it is prudent to administer all forms within a school and obtain school-level results by aggregating individual student scores across forms.

4. When school scores by form are to be reported and used, an estimate of the degree to which these scores will vary due to school-by-form interactions should be provided.

5. Any estimate of the possible variation in school scores is better than no estimate at all.

6. The magnitude of school-by-form interactions was found to vary by school size, presumably because school size is an indirect measure of other school differences. Therefore, it seems prudent to estimate school-by-form interactions separately for groups of schools that differ in size.

To summarize what was learned from this study, we can say that even when forms are equated, if they are short and not parallel, school scores can vary substantially across the forms. Equating does not fix this kind of variation. Making forms parallel and long would reduce this problem considerably. How parallel and how long must be addressed in future research. Until this is known, the best approach to ensuring that school scores can be compared both within and across test years is to make sure that the test forms on which any form-to-form comparisons are based cover the same content and are as similar as possible in terms of their psychometric characteristics.

REFERENCES

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Burket, G.R. (1996). *PARDUX* (Version 4.1). Monterey, CA: CTB/McGraw-Hill.

Cronbach, L.J., Bradburn, N. M., & Horvitz, D.G. (1994). *Report of the select committee: Sampling and statistical procedures used in the California Learning Assessment System.*

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.

Lane, S., Stone, C.A., Ankenmann, R. D., & Liu, M. (1993). Reliability and validity of a mathematics performance assessments. *International Journal of Educational Research, 21*, 247-262.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessment. *Journal of Educational Measurement, 30*, 215–232.

Yen, W.M. (1997a). The technical quality of performance assessment: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice, 16*, 5–15.

Yen, W.M. (1997b, March). *Measuring school performance: Is the percent of students reaching a standard the most accurate statistic?* Division D Invited Address presented at the Annual Meeting of the American Educational Research Association, Chicago.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

APPENDIX

Calculation of Statistics Used in the Analyses

Given that *j*th student in school *i* (y = 1 to Y) is administered Form f (f = 1 to F) and achieves a score of X_{yjf} on the form, the average score on this form is

$$M_{if} = \sum_{j=1}^{n_f} \frac{X_{yjf}}{n_{yf}} ,$$

where n_f refers to the number of students in the school who took the form. The within-school variance of the student scores on this form can be expressed as

$$Var(X_{yjf}) = \sum_{j=1}^{n_f} \frac{(X_{yjf} - M_{yf})^2}{n_{yf}} .$$

The within school variance in students' mean scores averaged across the forms is

$$Var\ (M_{yf.}) = \Sigma\ (M_{yf} - M_{y.})^2/F.$$

This variance can be summed across schools and averaged to find the average within school variance in mean scores across schools. The square root provides the SDM or the SDP:

$$SDM \quad or \quad SDP = \sqrt{(\sum_{y=1}^{Y} Var(M_{yf}))/Y} \quad .$$

The within-school variance of students' scores on a form can be summed across forms within a school and averaged. This average can then be summed across all schools and again averaged to estimate the average within-school variance of student scores on a form. This estimate can be divided by N../F, where

$$N.. = \sum_{y=1}^{Y} \sum_{f=1}^{F} n_{yf} ,$$

the average number of students per form tested in the schools. The square root of this result provides the SDM_s or the SDP_s.

TABLE 1: SKILLS MEASURED BY THE PA ITEMS IN EACH FORM

Grade	Form	Item No.	Skill
4	A	1	Collect/organize/display information; formulate questions and interpret information
		2	Use models to express the relationship between the dimensions and area of a rectangle
		3	Use one or more concrete models to represent fractions
	B	1	Same skills as Form A, Item 1
		2	Demonstrate or describe patterns in geometry, data collection, and arithmetic operations
		3	Model and explain division in a variety of ways
		4	Same skills as Form A, Item 2
	C	1	Use appropriate language and notation to describe and compare amounts of money
		2	Solve addition and subtraction problems involving multi-digit numbers
		3	Use models and pictures to show value of and to compare decimal numbers having tenths and hundredths places
8	A	1	Use patterns and algebraic methods to solve problems
		2	Collect, plot, and interpret data for two variables
		3	Analyze data using tables, graphs, rules, and equations
	B	1	Create, describe, and analyze geometric and numerical patterns
		2	Compute and interpret the mean and compare it to the median; use statistics to solve problems
		3	Find the surface area and volume of pyramids, prism, cylinders and cones
		4	Apply statistics to solve problems
	C	1	Use and explain definitions and laws related to exponents
		2	Analyze data using tables, graphs, rules, and equations; generate ordered pairs and graph a linear equation
		3	Estimate an answer, then solve complex problems and evaluate accuracy of results.

TABLE 2: SCHOOL SIZE CATEGORIES USED IN EACH GRADE

Grade	Size Category	Category Definition	Number of Schools	Percentage of Schools
4	Small	≤ 60 students	442	40
	Medium	61–120 students	542	48
	Large	> 120 students	134	12
	TOTAL		1118	
8				
	Small	≤ 120 students	242	45
	Medium	121–240 students	209	39
	Large	> 240 students	88	16
	TOTAL		539	

TABLE 3: DESCRIPTIVE RAW SCORE TEST STATISTICS

Grade	Form	No. of Items	Raw Score Mean	Raw Score SD	Maximum Possible Score	OPM	Stratified Alpha
4	A	3	3.2	1.9	9	36	0.56
	B	4	4.6	2.7	12	38	0.63
	C	3	3.1	2.2	9	34	0.52
8	A	3	2.0	1.8	9	22	0.55
	B	4	1.6	1.6	12	13	0.45
	C	3	2.0	1.9	9	22	0.55

TABLE 4: RESULTS OF SCALING MC AND PA ITEMS

Grade	Form	N of Items MC	N of Items PA	N of Items with Poor Fit MC	N of Items with Poor Fit PA	N of Item Pairs with $Q_3 \geq 2.00$ MC-MC	MC-PA	PA-PA	N of Non-converging Items MC	N of Non-converging Items PA
4	A	80	3		1	1				
	B	80	4	1		2				
	C	80	3			2				
8	A	80	3	2					2	
	B	80	4						2	
	C	80	3	1					2	

TABLE 5: SCALE SCORE STATISTICS DESCRIBING STUDENT-LEVEL PERFORMANCE FOR ALL STUDENTS IN THE STATE

Grade	Form	Number of Students	Scale Score			Percentage at	
			Mean	SD	Median	300	700
4	A	28,821	492	89	499	6	1
	B	28,103	491	88	495	5	2
	C	27,543	490	97	505	14	2
8	A	26,935	464	112	487	27	1
	B	26,404	462	112	481	27	0
	C	25,844	464	110	503	27	0

TABLE 6: STATISTICS* DESCRIBING SCHOOL-LEVEL SCALE SCORES FOR ALL SCHOOLS

Grade	Form	No. of Schools	Average School Size	School Means		Average SDWS	(A) Average SDM	(B) SDM_e	$[(A^2-B^2)/A^2)]^*$ x100
				Average	SD				
4	A	1118	76	489	38	84	21	17	34%
	B			489	39				
	C			493	38				
8	A	539	147	459	48	103	28	15	71%
	B			455	46				
	C			457	47				

*SDWS = standard deviation of student scores on a form within a school
SDM = standard deviation of form means within a school
SDM_e = estimated standard deviation of a form mean due to sampling

TABLE 7: STATISTICS* DESCRIBING SCHOOL-LEVEL SCALE SCORES BY SCHOOL SIZE

Size Category/ Grade	Form	No. of Schools	Average School Size	School Means Average	School Means SD	Average SDWS	(A) Average SDM	(B) SDM_e	$[(A^2-B^2)/A^2)]$ x100
Small									
4	A	442	41	483	43	83	26	23	22%
	B			485	44				
	C			490	42				
8	A	242	59	452	56	100	34	23	54%
	B			445	54				
	C			447	58				
Medium									
4	A	542	85	492	34	85	19	16	29%
	B			490	36				
	C			495	35				
8	A	209	185	462	38	105	22	13	65%
	B			463	35				
	C			463	33				
Large									
4	A	134	151	499	33	85	15	12	36%
	B			497	34				
	C			499	31				
8	A	88	299	470	44	105	21	11	73%
	B			466	39				
	C			470	38				

*SDWS = standard deviation of student scores on a form within a school
SDM = standard deviation of form means within a school

TABLE 8: MEAN SCALE SCORES BY FORM FOR THE GRADE 4 SCHOOLS WITH THE LARGEST SDM_yS, BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of Means	SDM_y
Small	1	24	390	475	580	190	78
	2	18	471	423	603	180	76
	3	20	580	463	414	166	69
	4	44	349	435	509	160	66
	5	24	391	521	524	133	62
Medium	1	86	488	496	597	109	50
	2	66	437	486	556	119	49
	3	93	528	441	424	104	45
	4	67	416	497	521	105	45
	5	61	439	508	546	107	44
Large	1	179	487	553	511	66	27
	2	138	511	575	554	64	27
	3	122	517	469	456	61	26
	4	126	478	494	539	61	26
	5	131	509	513	458	55	25

TABLE 9: MEAN SCALE SCORES BY FORM FOR THE GRADE 8 SCHOOLS WITH THE LARGEST SDM_yS, BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of Means	SDM_y
Small	1	19	660	397	615	263	115
	2	9	367	525	534	167	76
	3	14	337	472	300	172	74
	4	13	514	336	430	178	73
	5	17	548	380	511	168	72
Medium	1	204	497	365	463	132	56
	2	179	380	499	407	119	51
	3	234	528	416	520	112	51
	4	144	426	524	535	109	49
	5	217	387	492	420	105	44
Large	1	244	537	401	476	136	56
	2	286	460	391	487	96	40
	3	281	368	458	447	90	40
	4	329	495	421	508	87	38
	5	309	367	445	449	82	38

TABLE 10: MEAN SCALE SCORES BY FORM FOR THE GRADE 4 SCHOOLS WITH THE SMALLEST SDM_yS, BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of Means	SDM_y
Small	1	34	478	481	482	4	1
	2	32	490	486	487	4	2
	3	26	521	524	520	4	2
	4	42	497	493	497	4	2
	5	60	464	465	469	5	2
Medium	1	77	504	505	504	1	0
	2	112	502	500	503	3	1
	3	109	508	504	506	4	2
	4	63	490	492	487	5	2
	5	109	521	525	526	5	2
Large	1	144	483	485	485	2	1
	2	164	494	494	492	2	1
	3	151	513	512	509	4	2
	4	200	494	497	492	5	2
	5	121	506	503	510	7	3

TABLE 11: MEAN SCALE SCORES BY FORM FOR THE GRADE 8 SCHOOLS WITH THE SMALLEST SDM_yS, BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of Means	SDM_y
Small	1	113	455	456	459	3	2
	2	17	454	457	462	8	3
	3	30	518	529	523	11	4
	4	42	509	503	492	17	7
	5	111	478	460	464	18	8
Medium	1	193	460	456	454	6	2
	2	177	437	441	444	7	3
	3	173	461	467	469	8	3
	4	184	469	470	476	7	3
	5	134	522	530	529	8	4
Large	1	274	463	466	464	3	1
	2	304	531	524	533	9	4
	3	273	499	500	507	8	4
	4	245	445	436	437	9	4
	5	307	499	501	491	10	5

TABLE 12: STUDENT PAACS BY FORM FOR ALL STUDENTS IN THE STATE

Grade	Form	N	PAAC
4	A	28,821	61%
	B	28,103	61%
	C	27,543	58%
8	A	26,935	55%
	B	26,404	66%
	C	25,844	52%

TABLE 13: STATISTICS* DESCRIBING SCHOOL PAACS FOR ALL SCHOOLS

		N of	Average	School PAACs		Average		$[(A^2-B^2)/A^2)]$
Grade	Form	Schools	School Size	Average	SD	SDP	SDP_e	x100
4	A	1118	76	60%	18	11	9	33%
	B			60%	19			
	C			57%	19			
8	A	539	147	53%	19	16	7	81%
	B			64%	20			
	C			49%	19			

*SDP = standard deviation of form PAACs within a school
SDP_e = estimated standard deviation of a form PAAC due to sampling

TABLE 14: STATISTICS* DESCRIBING SCHOOL PAACS BY SCHOOL SIZE CATEGORY

Size Category/ Grade	Form	N of Schools	Average School Size	School PAACs Average	School PAACs SD	(A) Average SDP	(B) SDP_s	$[(A^2-B^2)/A^2)]$ x100
Small								
4	A	442	41	57%	21	14	13	14%
	B			58%	22			
	C			56%	21			
8	A	242	59	50%	23	16	10	61%
	B			60%	24			
	C			46%	23			
Medium								
4	A	542	85	61%	16	10	9	19%
	B			60%	17			
	C			58%	18			
8	A	209	185	54%	15	12	6	75%
	B			67%	14			
	C			51%	14			
Large								
4	A	134	151	64%	16	8	7	23%
	B			64%	14			
	C			59%	15			
8	A	88	299	57%	17	11	5	79%
	B			68%	16			
	C			54%	15			

*SDP = standard deviation of form PAACs within a school
SDP_s = estimated standard deviation of a form PAAC due to sampling

TABLE 15: PAACS BY FORM FOR THE GRADE 4 SCHOOLS WITH THE LARGEST SDP_yS BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of PAACs	SDP_y
Small	1	20	72%	86%	0%	86	38
	2	24	12%	58%	100%	88	36
	3	44	0%	54%	86%	86	35
	4	13	20%	50%	100%	80	33
	5	9	34%	100%	100%	66	31
Medium	1	74	93%	80%	38%	55	23
	2	71	66%	14%	61%	52	23
	3	68	81%	73%	28%	53	23
	4	97	32%	33%	81%	49	23
	5	86	45%	59%	97%	52	22
Large	1	143	82%	81%	39%	43	20
	2	154	88%	65%	44%	44	18
	3	123	54%	66%	26%	40	17
	4	124	30%	68%	40%	38	16
	5	140	62%	62%	30%	32	15

TABLE 16: PAACS BY FORM FOR THE GRADE 8 SCHOOLS WITH THE LARGEST SDP_yS BY SCHOOL SIZE CATEGORY

Category	School	School Size	A	B	C	Range of PAACs	SDP_y
Small	1	39	8%	86%	0%	86	39
	2	20	13%	100%	60%	87	36
	3	17	17%	100%	80%	83	35
	4	19	100%	29%	100%	71	33
	5	53	94%	62%	16%	78	32
Medium	1	123	28%	89%	11%	78	33
	2	179	29%	88%	27%	61	28
	3	124	26%	82%	20%	62	28
	4	132	83%	96%	35%	61	26
	5	152	33%	90%	70%	57	24
Large	1	337	37%	88%	49%	51	22
	2	281	18%	70%	46%	52	21
	3	309	14%	64%	45%	50	21
	4	276	50%	96%	55%	46	21
	5	309	62%	90%	40%	50	20

TABLE 17: CALCULATION OF STANDARD ERRORS AND CONFIDENCE INTERVALS FOR SCHOOL MEAN SCALE SCORES

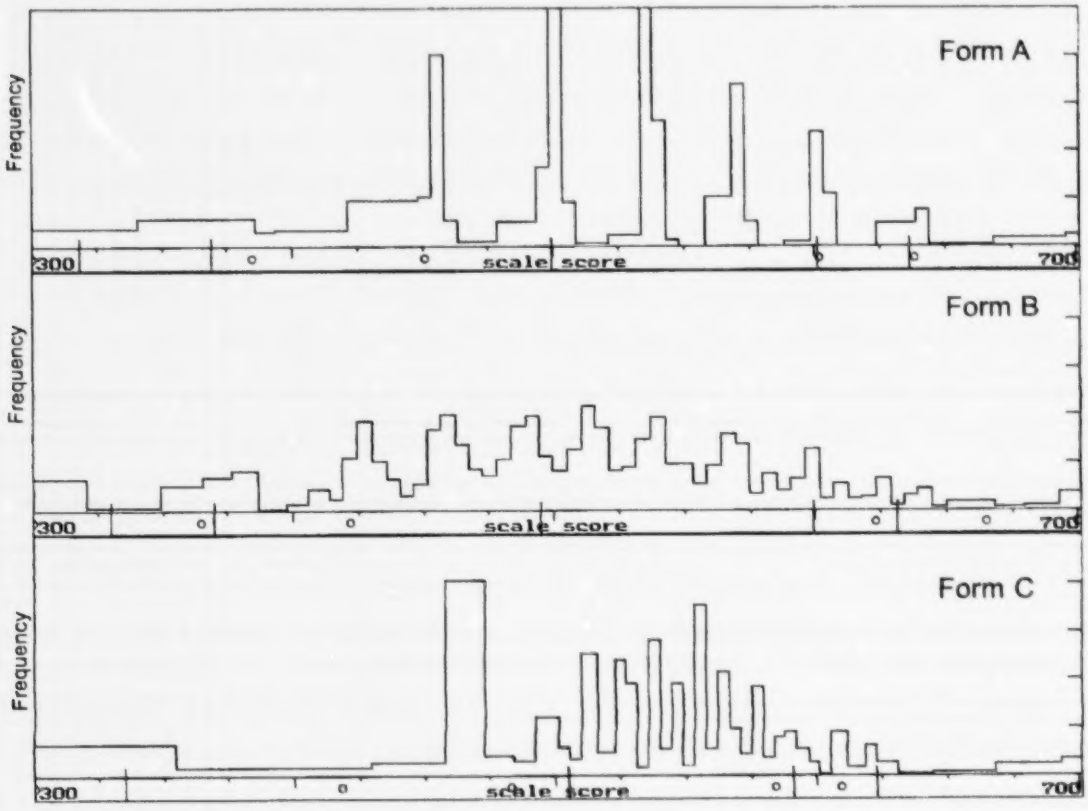
Grade	N of Schools	Size	Average SDM	SE(SS)= Average SDM/ $\sqrt{\text{N of Forms}}$	1.65*SE(SS)	Width of Confidence Interval
4	1118	All	21	12	20	40
	442	Small	26	15	25	50
	542	Medium	19	11	18	36
	134	Large	15	9	14	28
8	539	All	28	16	27	54
	242	Small	34	20	32	64
	209	Medium	22	13	21	42
	88	Large	21	12	20	40

TABLE 18: CALCULATION OF STANDARD ERRORS AND CONFIDENCE INTERVALS FOR SCHOOL PAACS

Grade	N of Schools	Size	Average SDP	SE (PAAC) = Average SDP/ $\sqrt{\text{N of Forms}}$	1.65*SE(PAAC)	Width of Confidence Interval
4	1118	All	11	6	10	21
	442	Small	14	8	13	26
	542	Medium	10	6	10	19
	134	Large	8	5	8	16
8	539	All	16	9	15	30
	242	Small	16	9	15	30
	209	Medium	12	7	11	22
	88	Large	11	6	10	20

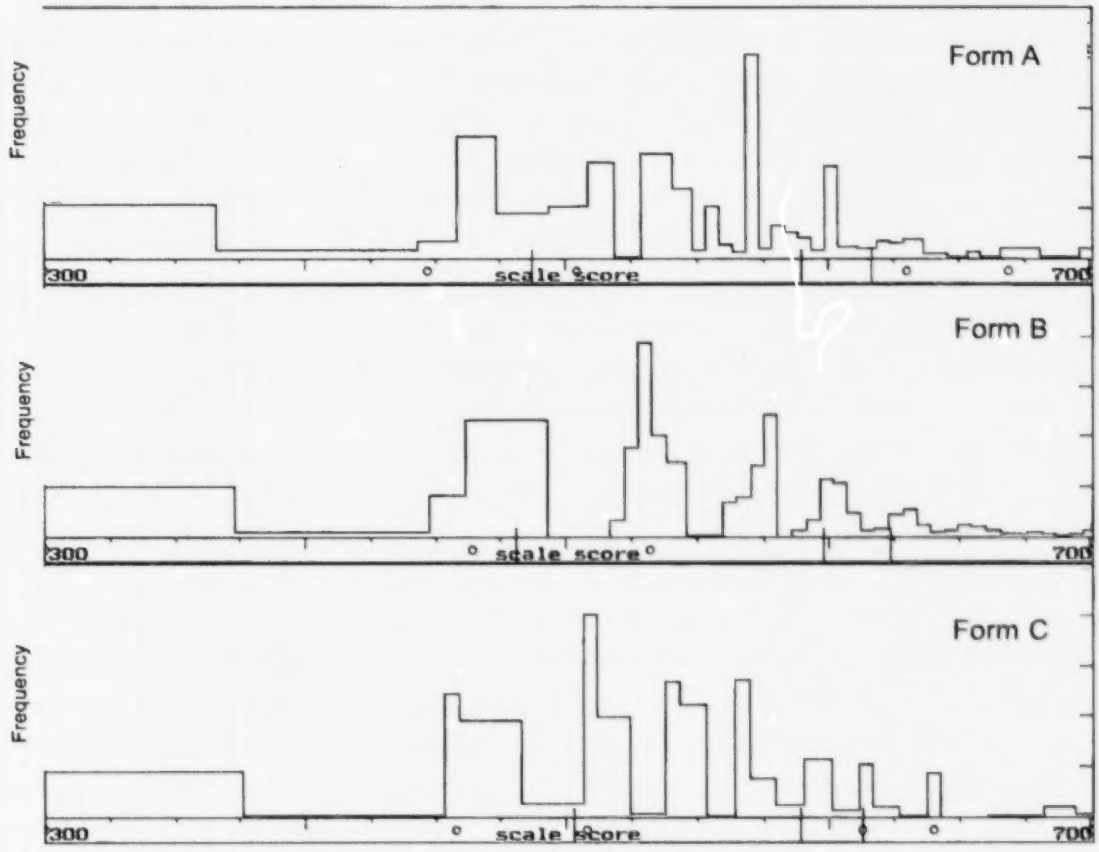
FIGURES

FIGURE 1: FREQUENCY DISTRIBUTIONS OF THE GRADE 4 EQUATED SCORES FOR FORMS A, B, AND C.



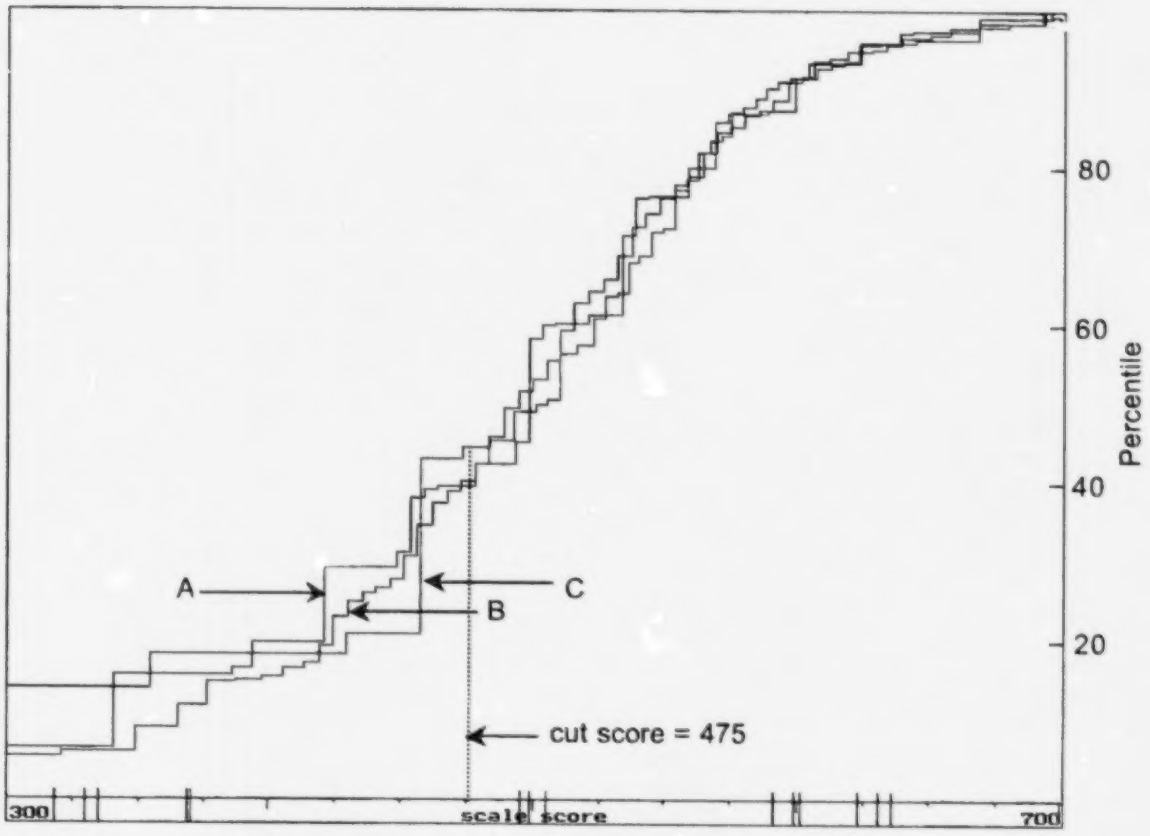
BEST COPY AVAILABLE

FIGURE 2: FREQUENCY DISTRIBUTIONS OF THE GRADE 8 EQUATED SCORES FOR FORMS A, B, AND C.



BEST COPY AVAILABLE

FIGURE 3: CUMULATIVE FREQUENCY DISTRIBUTION OF GRADE 4 EQUATED SCORES FOR FORMS A, B, AND C.



BEST COPY AVAILABLE

FIGURE 4: CUMULATIVE FREQUENCY DISTRIBUTION OF THE GRADE 8 EQUATED SCORES FOR FORMS A, B, AND C.

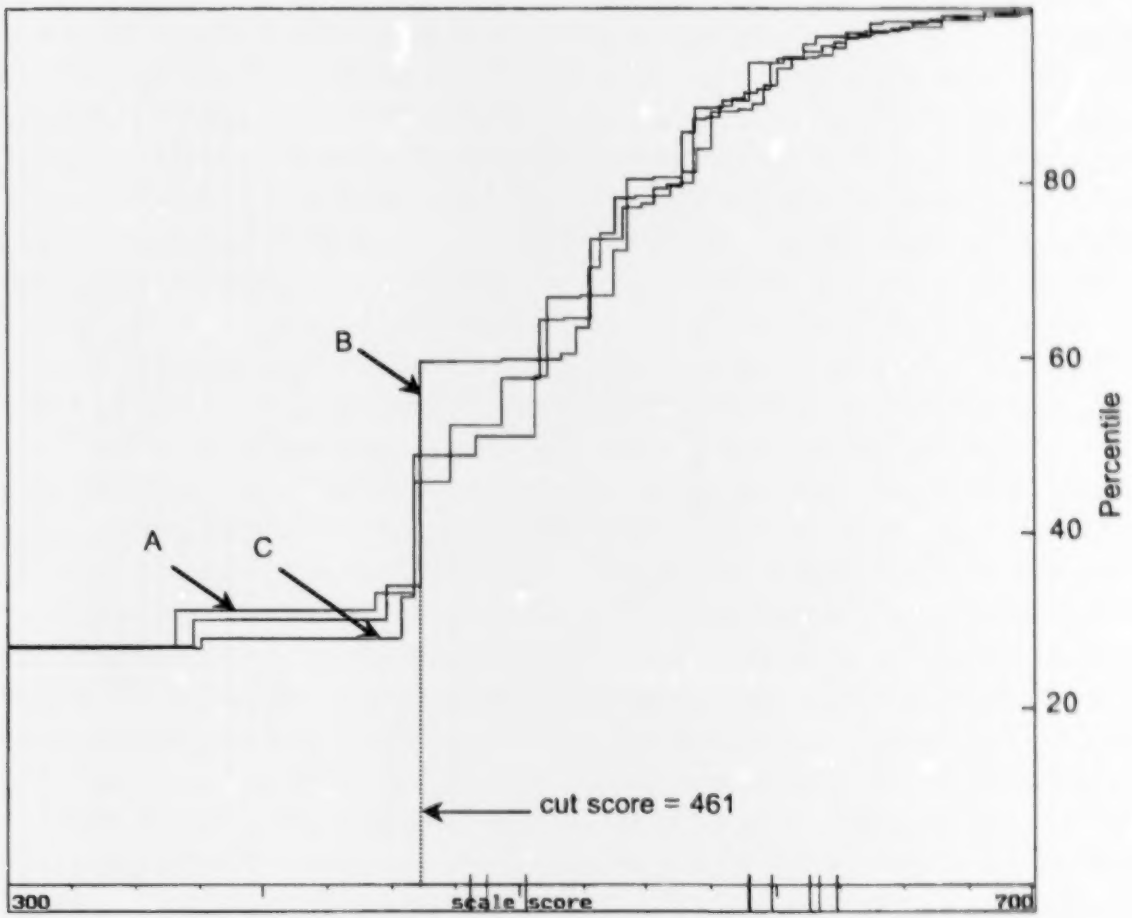


FIGURE 5A: STANDARD ERROR CURVES FOR THE GRADE 4 EQUATED FORMS A, B, AND C.

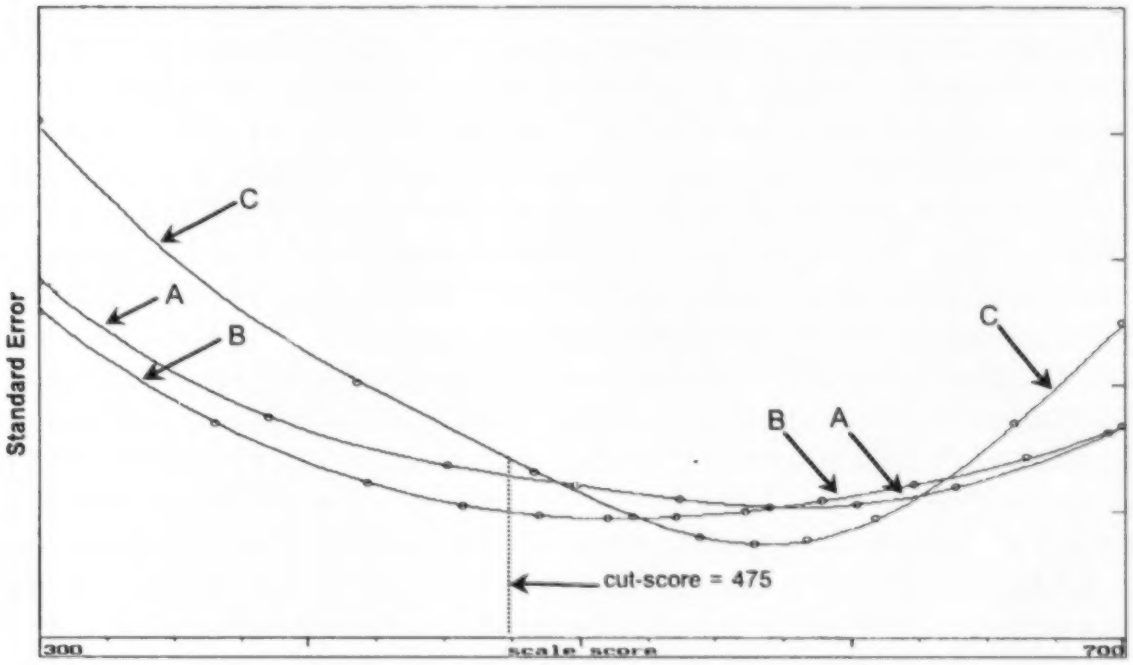


FIGURE 5b: STANDARD ERROR CURVES FOR THE GRADE 8 EQUATED FORMS A, B, AND C.

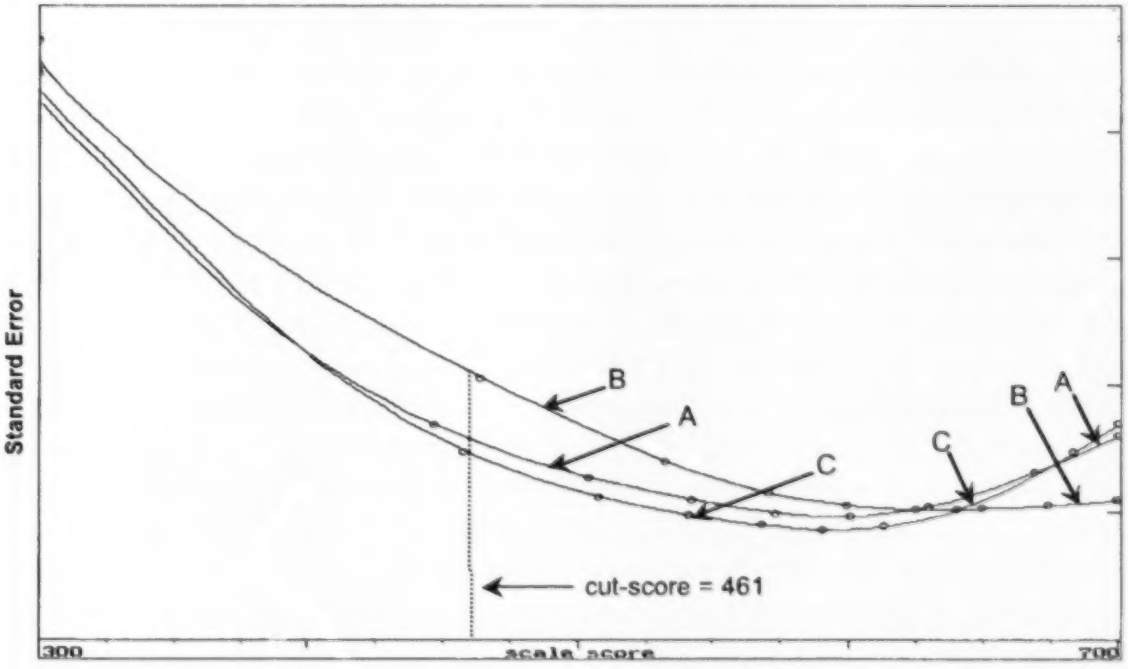


FIGURE 6a: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM_yS IN THE GRADE 4 SMALL SCHOOL CATEGORY

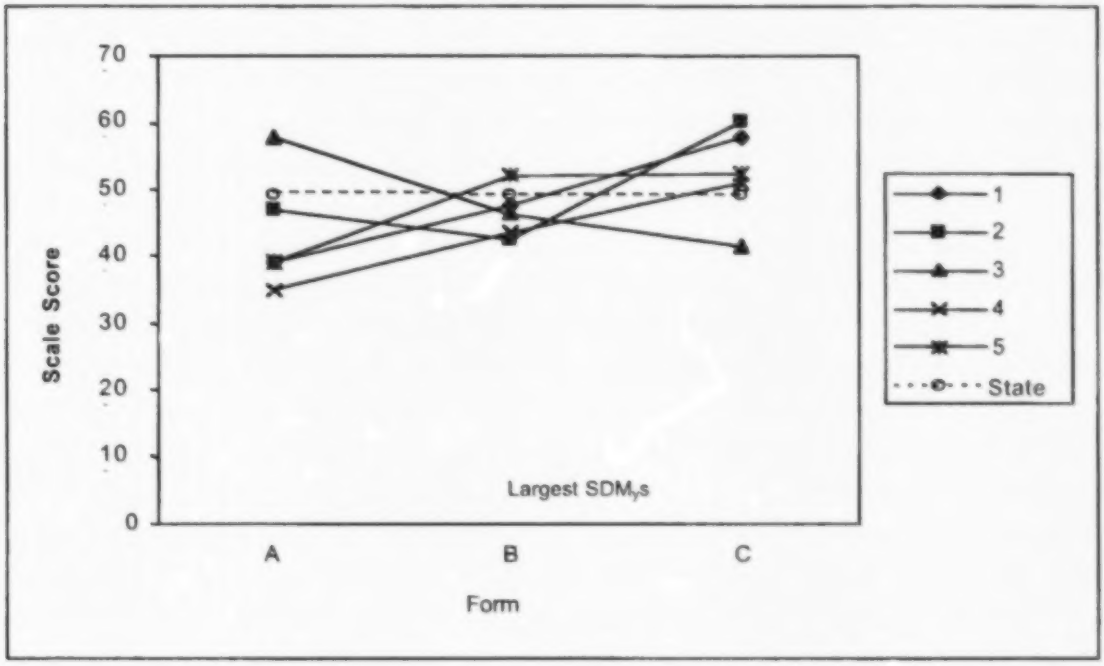


FIGURE 6b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM_yS IN THE GRADE 4 SMALL SCHOOL CATEGORY

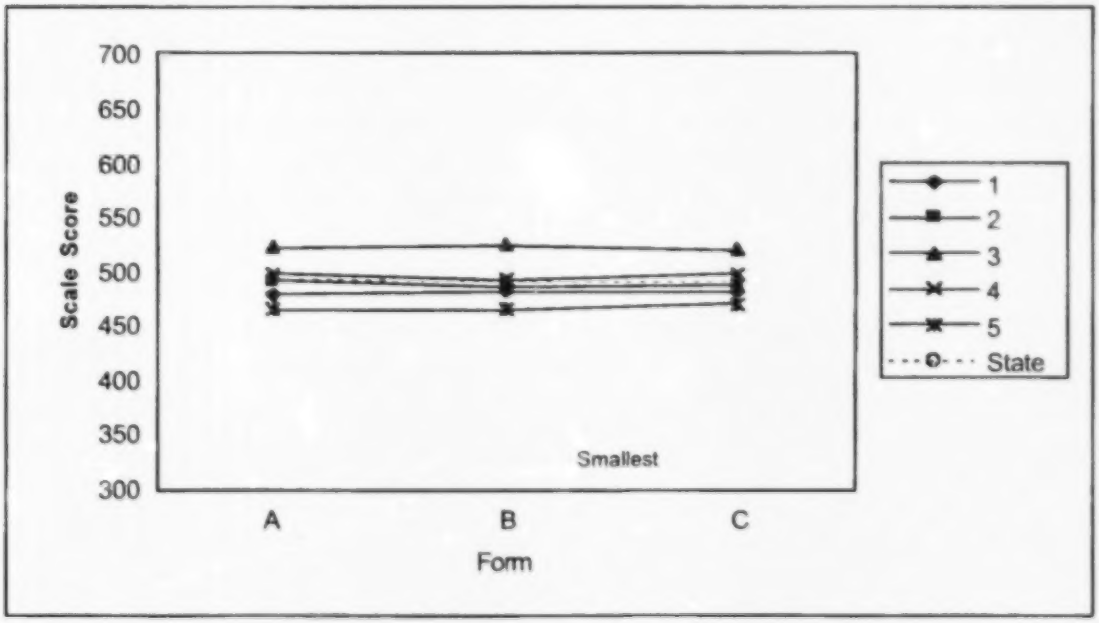


FIGURE 7a: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM$_y$S IN THE GRADE 4 SMALL SCHOOL CATEGORY.

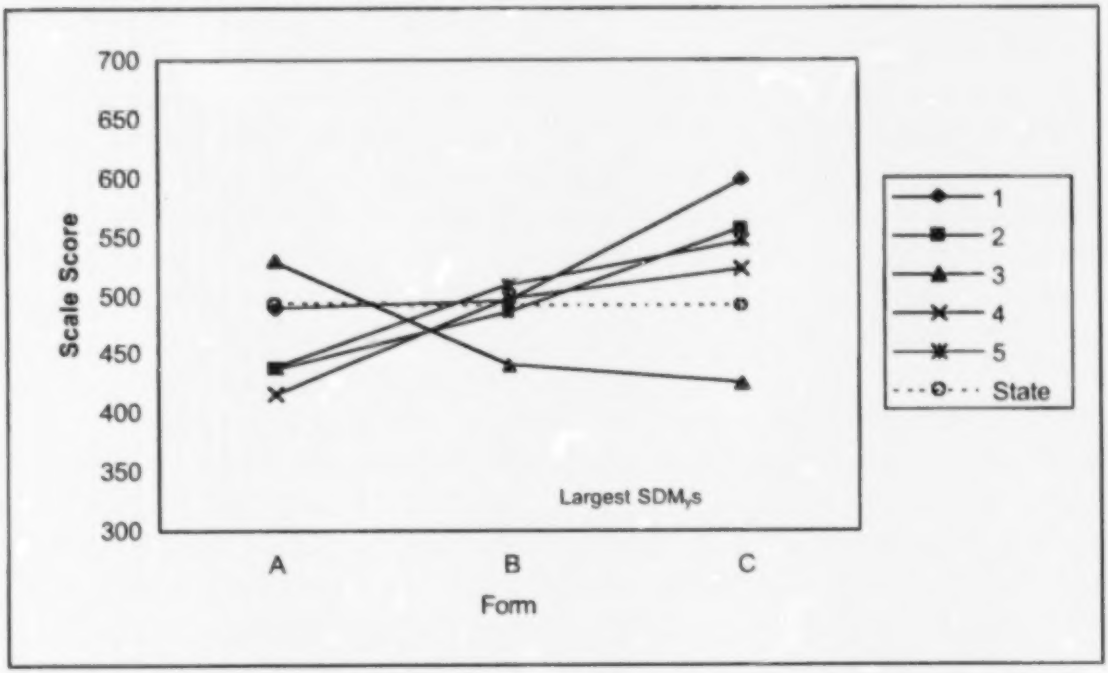


FIGURE 7b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM$_y$S IN THE GRADE 4 MEDIUM-SIZED SCHOOL CATEGORY

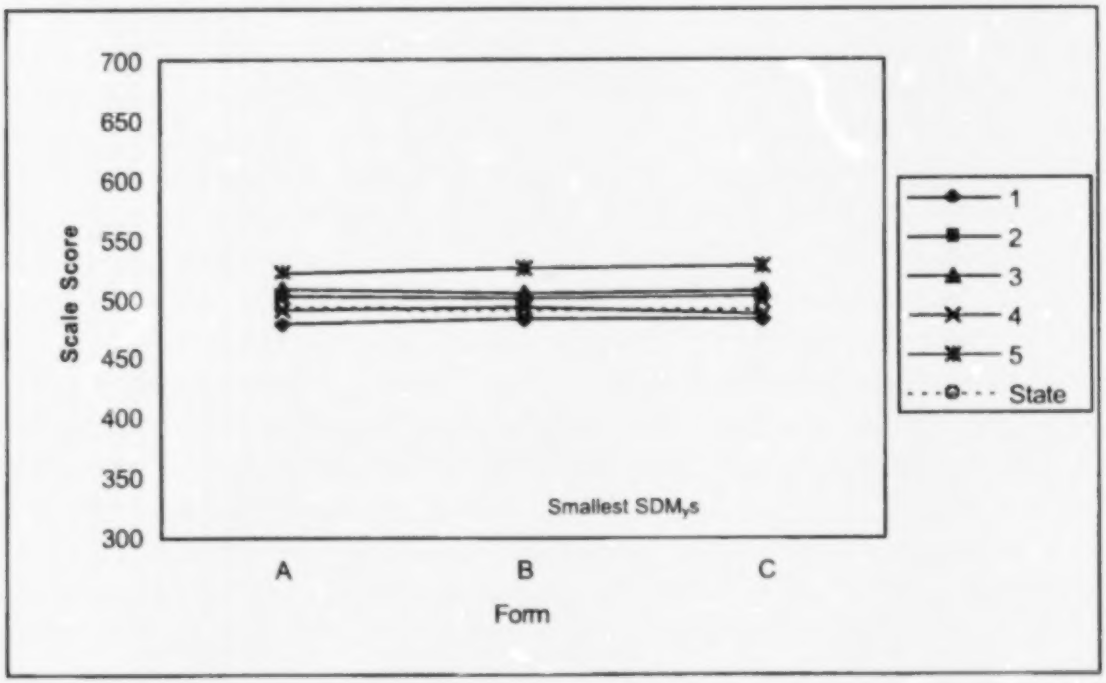


FIGURE 8A: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM_yS IN THE GRADE 4 LAGE SCHOOL CATEGORY

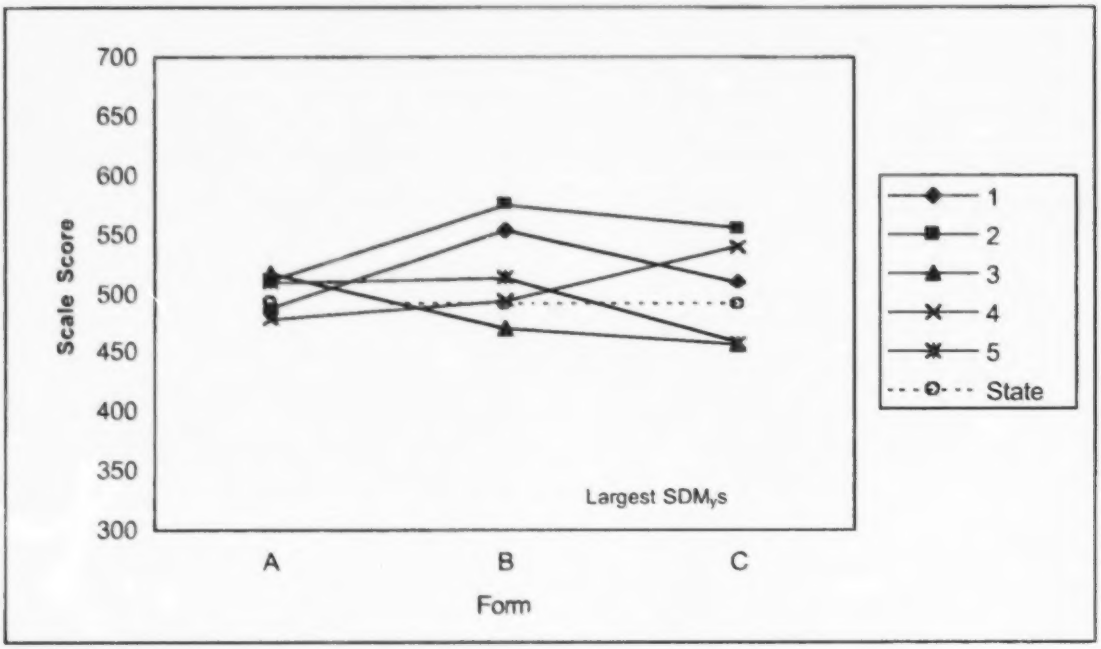


FIGURE 8b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM_yS IN THE GRADE 4 LARGE SCHOOL CATEGORY

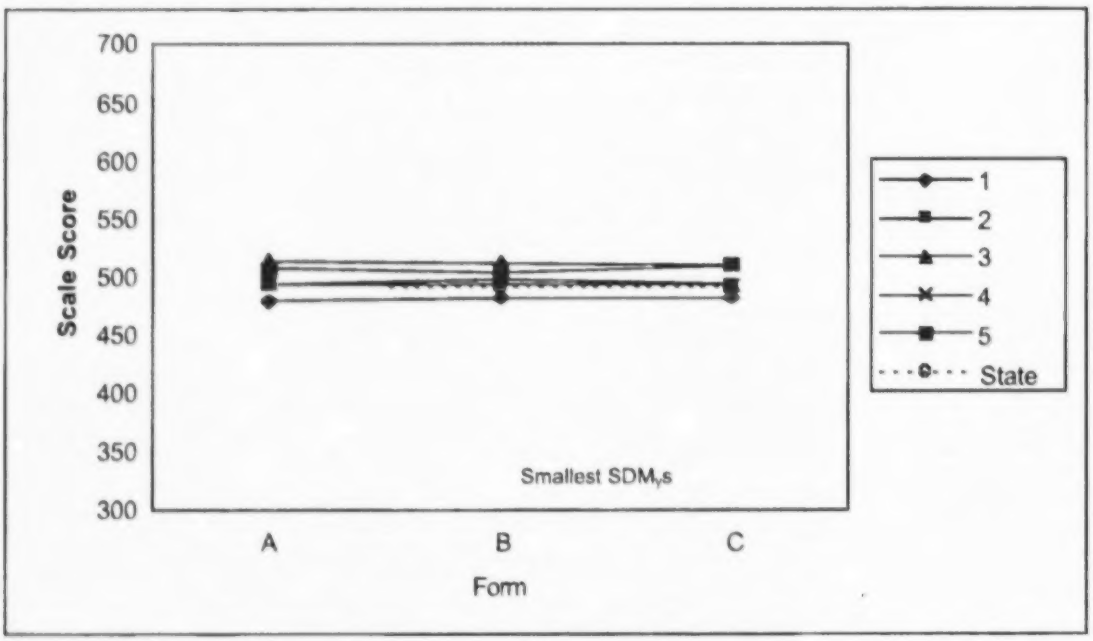


FIGURE 9a: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM_yS IN THE GRADE 8 SMALL SCHOOL CATEGORY.

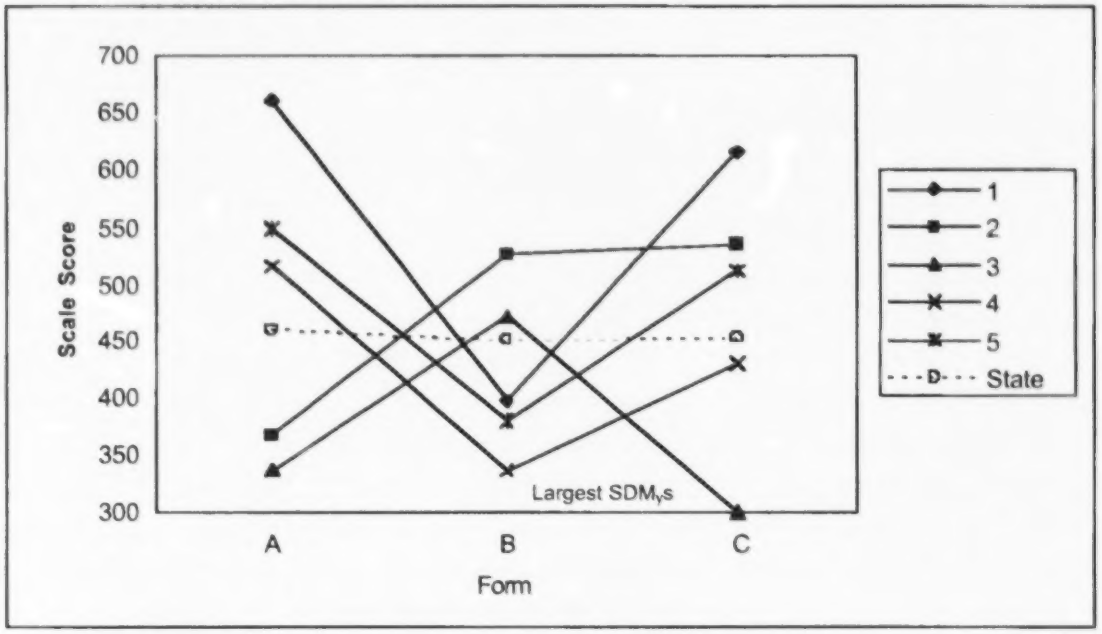


FIGURE 9b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM_yS IN THE GRADE 8 SMALL SCHOOL CATEGORY

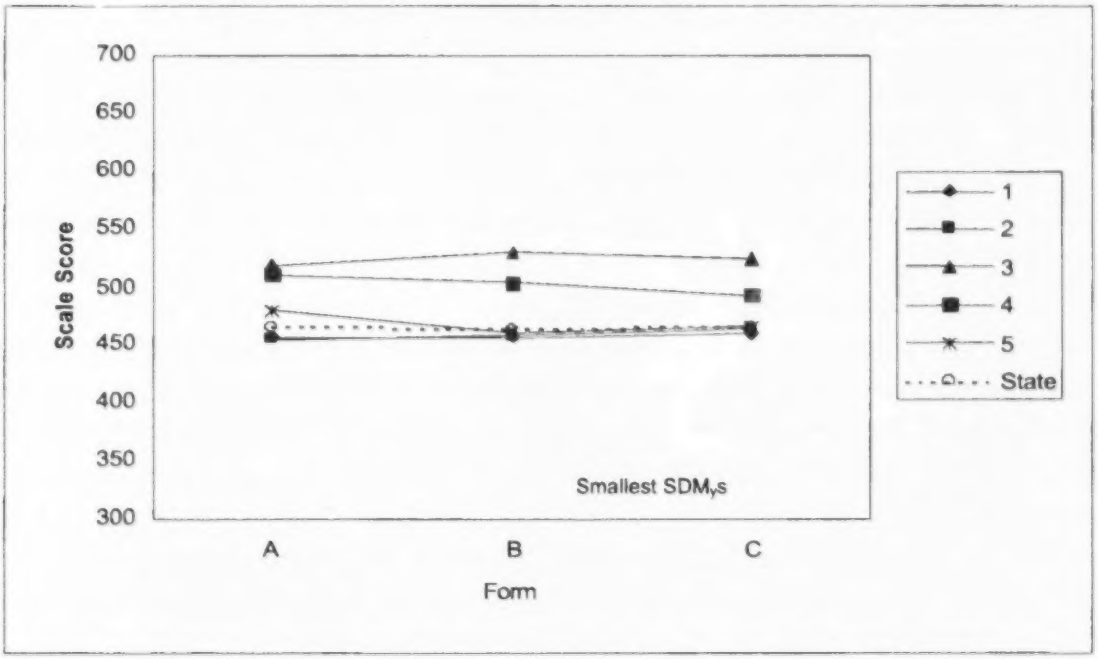


FIGURE 10a: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM_yS IN THE GRADE 8 MEDIUM-SIZED CATEGORY

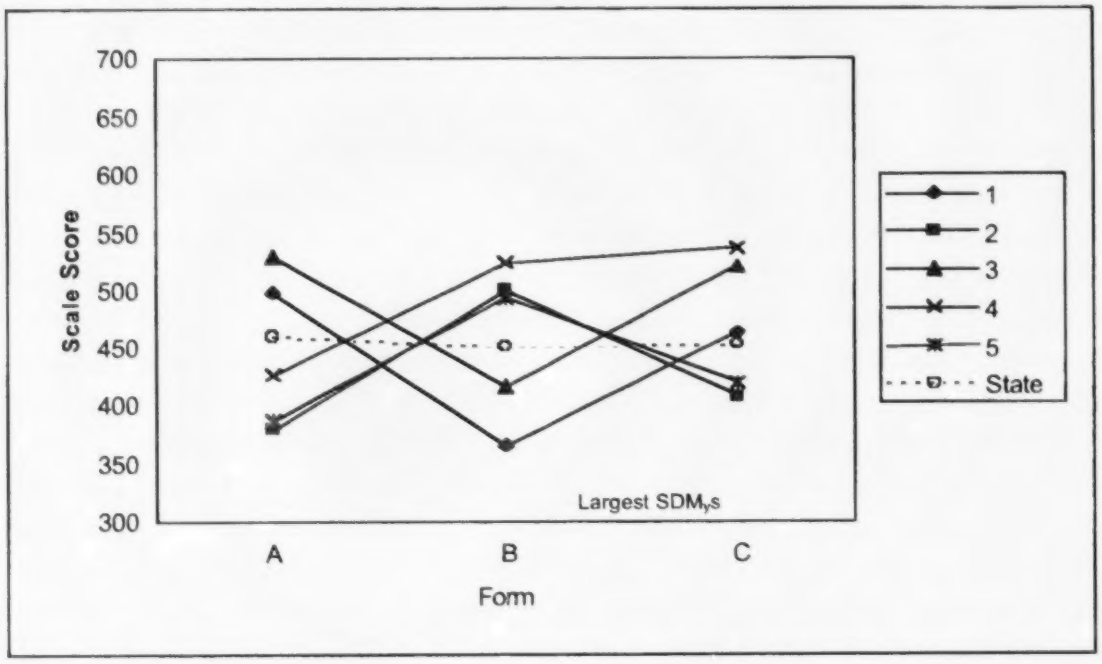


FIGURE 10b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM_yS IN THE GRADE 8 MEDIUM-SIZED SCHOOL CATEGORY

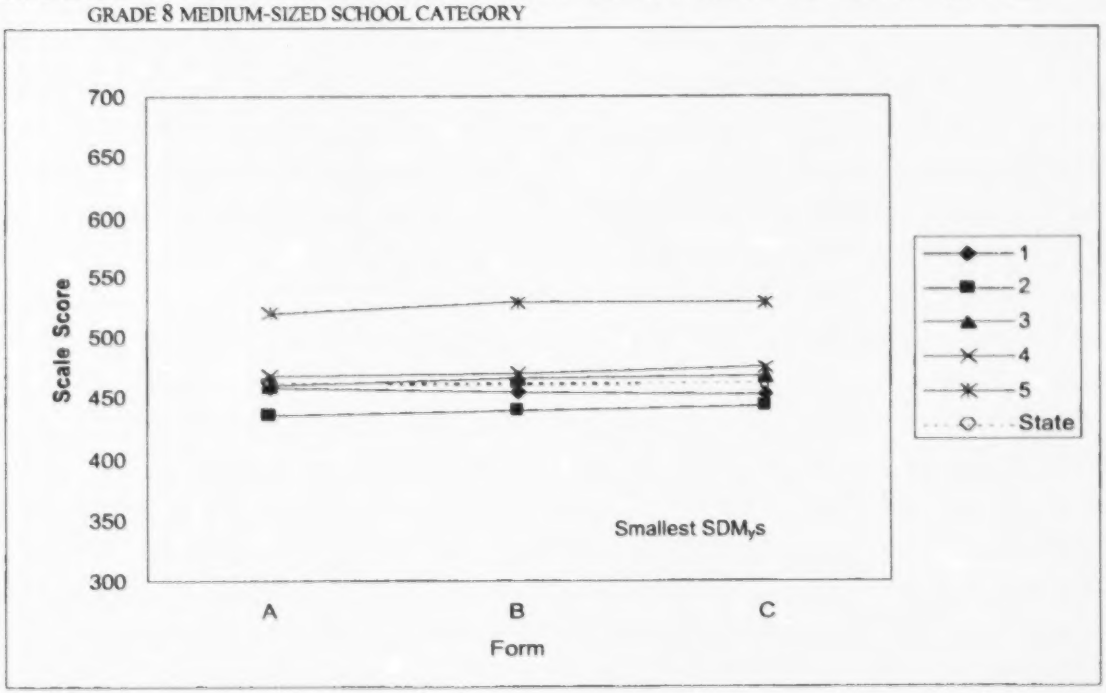


FIGURE 11A: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE LARGEST SDM_yS IN THE GRADE 8 SMALL SCHOOL CATEGORY.

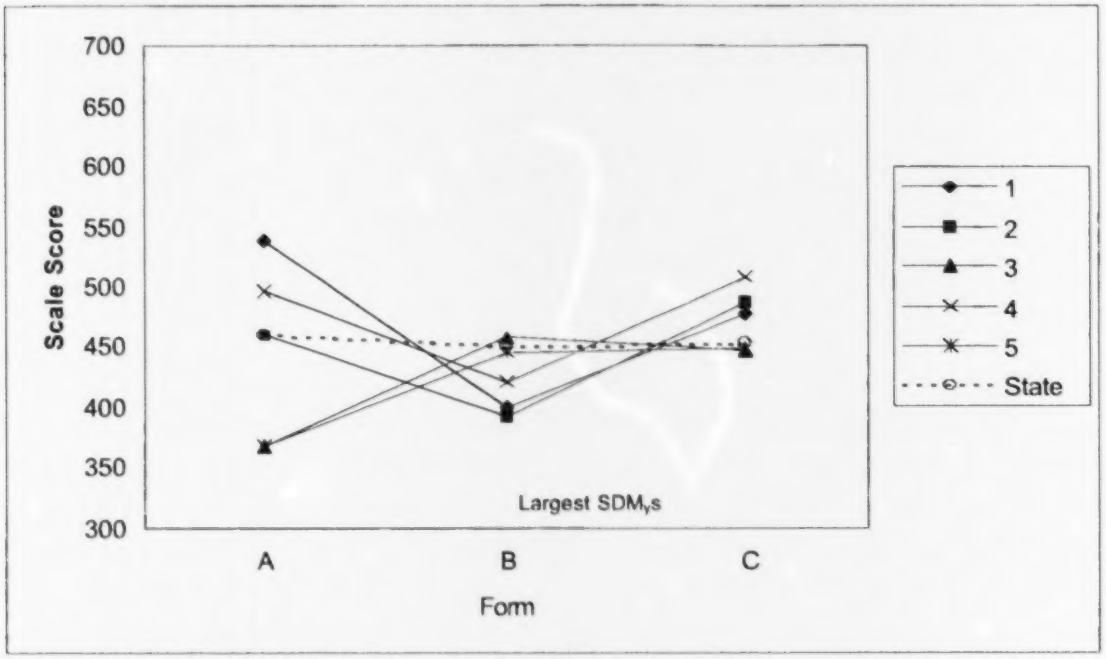


FIGURE 11b: MEAN SCALE SCORES BY FORM FOR THE FIVE SCHOOLS WITH THE SMALLEST SDM_yS IN THE GRADE 8 LARGE SCHOOL CATEGORY

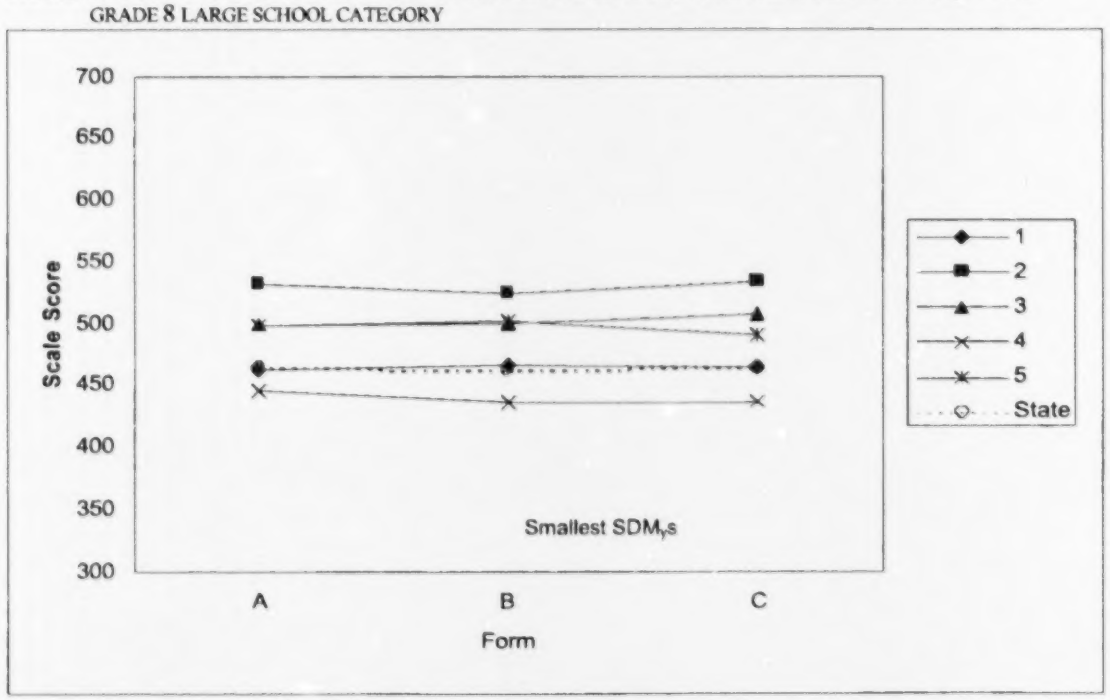
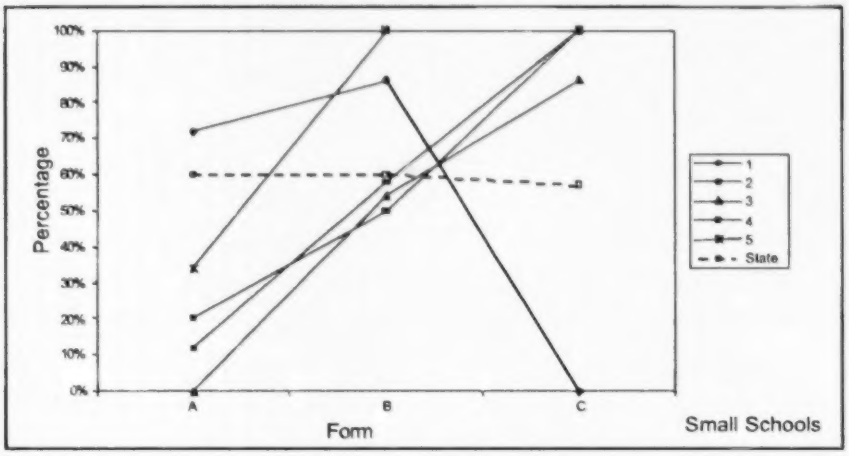
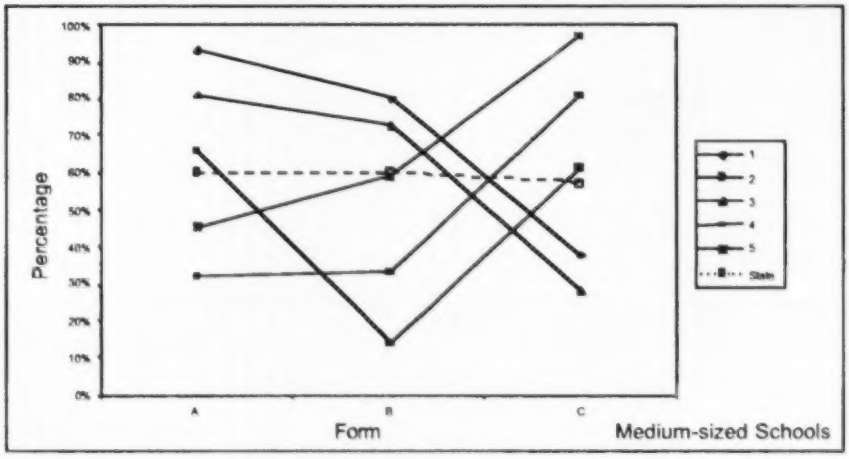


FIGURE 12: PAACs BY FORM FOR THE FIVE GRADE 4 SCHOOLS WITH THE LARGEST SDP,S





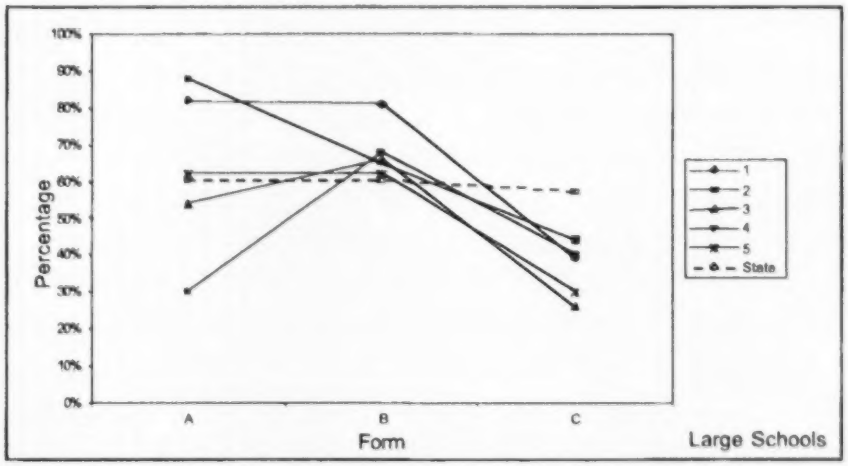
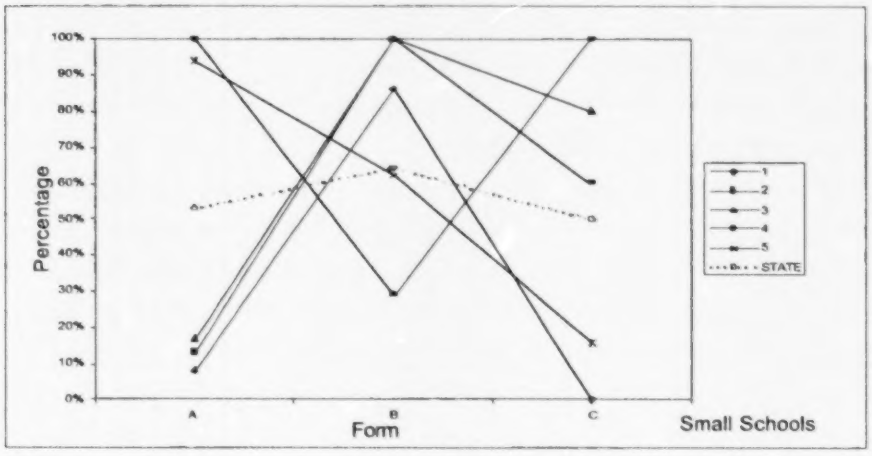
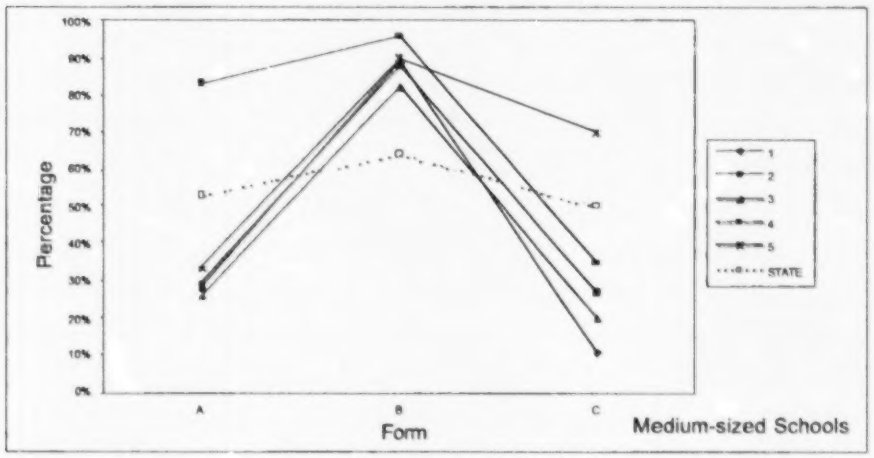


FIGURE 13: PAACS BY FORM FOR THE FIVE GRADE 8 SCHOOLS WITH THE LARGEST SDP,S





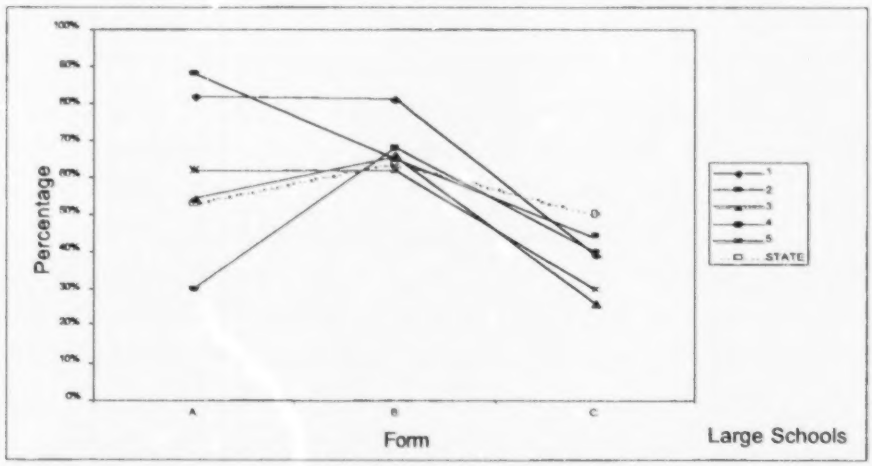


FIGURE 14: PERCENTAGE FREQUENCY DISTRIBUTIONS OF SDM_yS FOR GRADE 4 BY SCHOOL SIZE CATEGORY

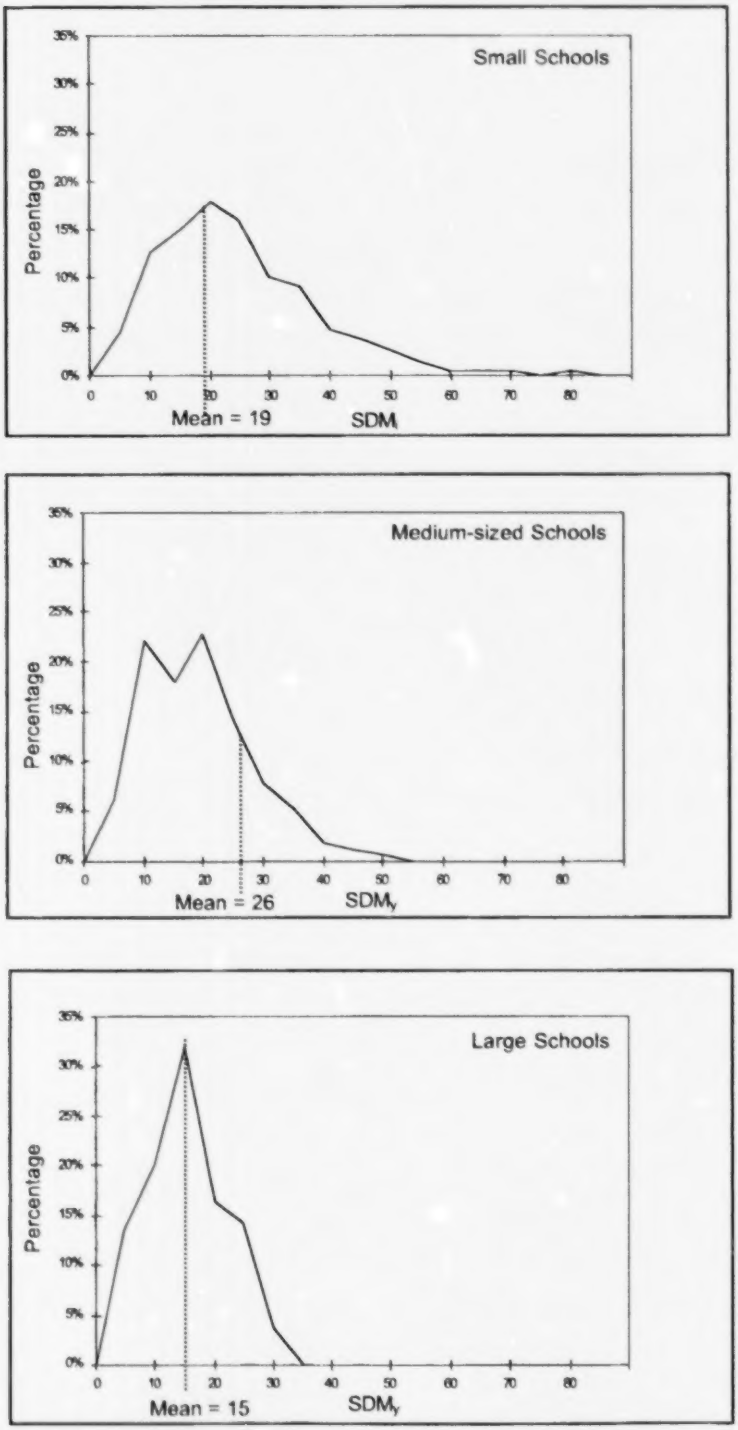


FIGURE 15: PERCENTAGE FREQUENCY DISTRIBUTIONS OF SDM_YS FOR GRADE 8 BY SCHOOL SIZE CATEGORY

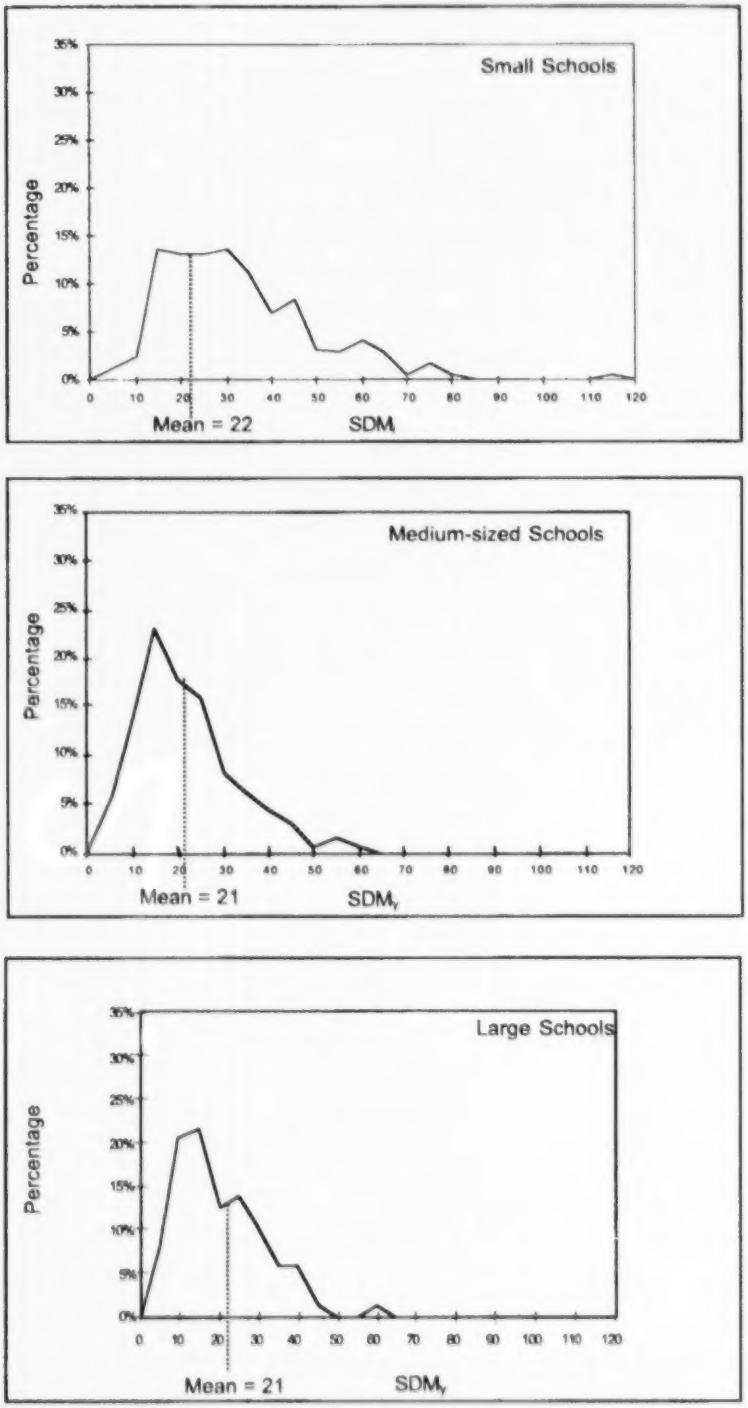
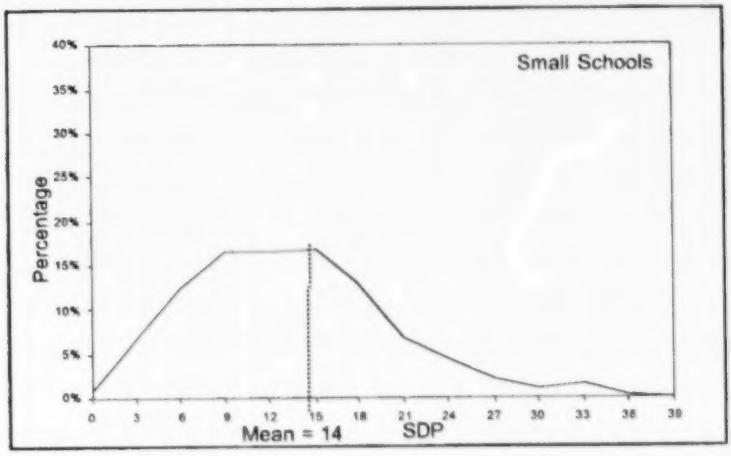
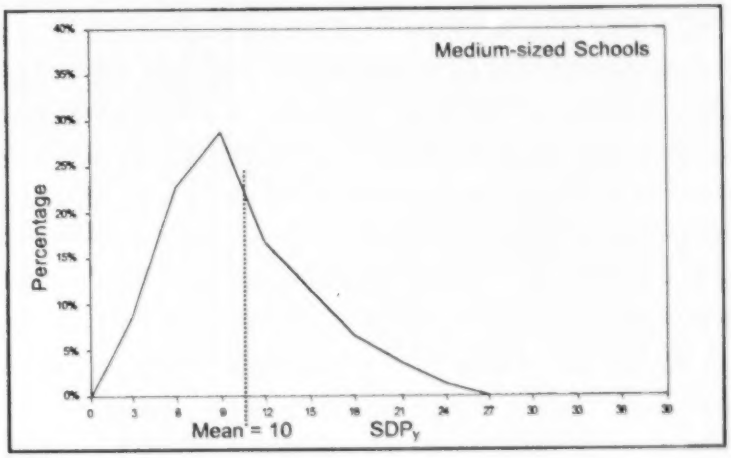


FIGURE 16: PERCENTAGE FREQUENCY DISTRIBUTIONS OF SDP_yS FOR GRADE 4 BY SCHOOL SIZE CATEGORY





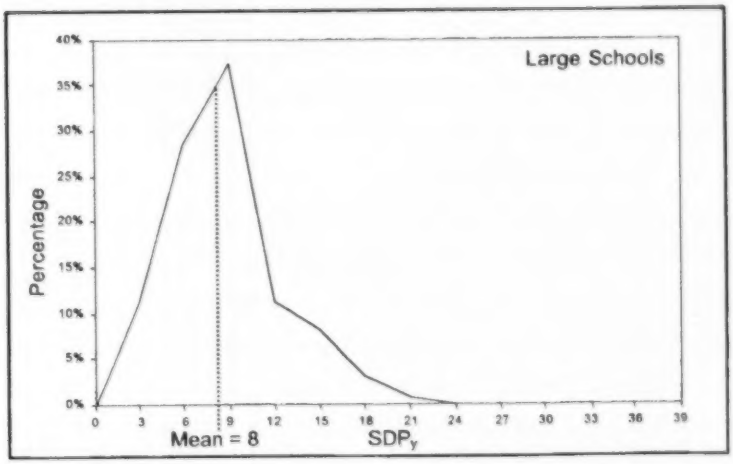
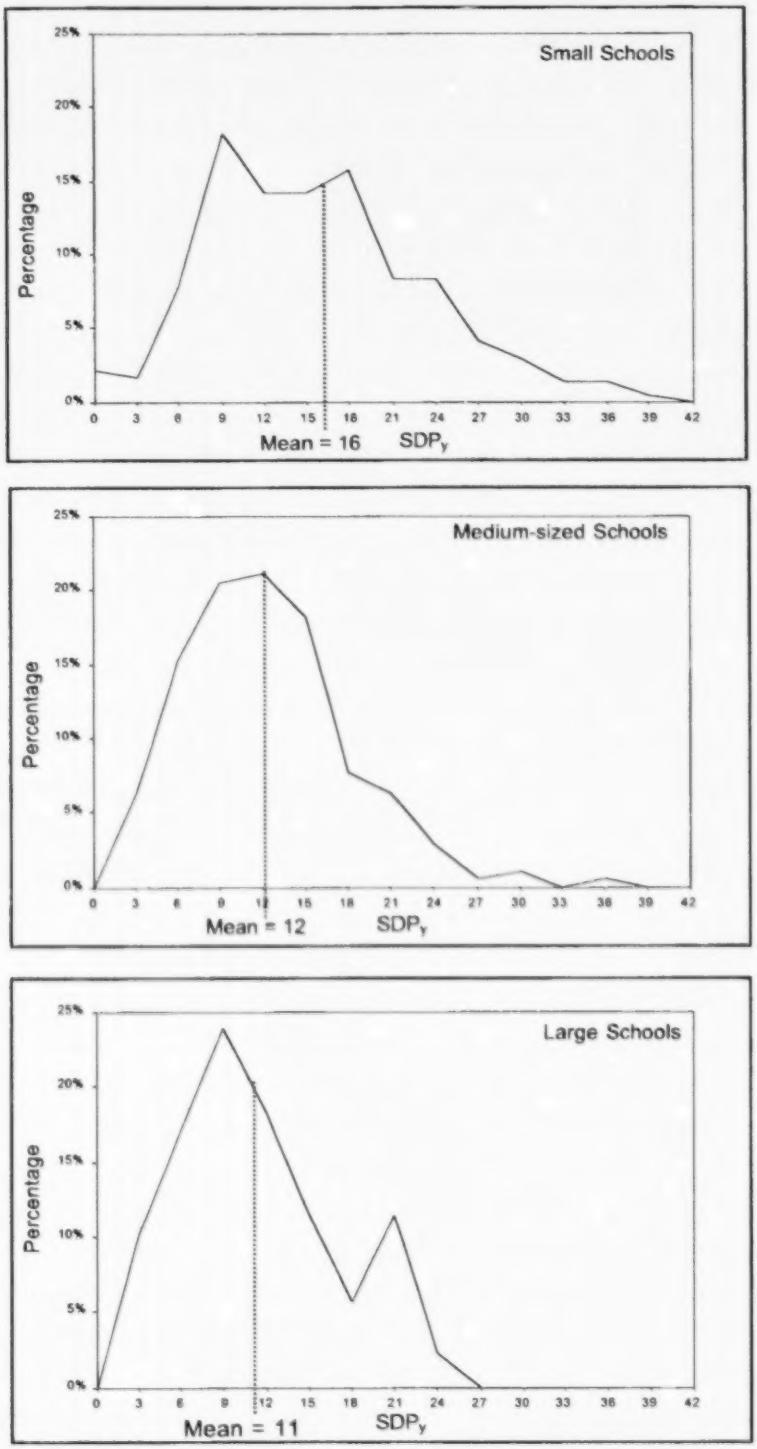


FIGURE 17: PERCENTAGE FREQUENCY DISTRIBUTIONS OF SDP_yS FOR GRADE 8 BY SCHOOL SIZE CATEGORY



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").